Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado

Facultad de Ingeniería de Sistemas e Informática

Unidad de Posgrado

"Método para recomendar factores de posicionamiento personalizados en el motor de búsqueda de Google"

TESIS

Para optar el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software

AUTOR

Richard Enrique INJANTE ORÉ

ASESOR

Dr. David Santos MAURICIO SÁNCHEZ

Lima, Perú

2020

Referencia bibliográfica

Injante, R. (2020). *Método para recomendar factores de posicionamiento personalizados en el motor de búsqueda de Google*. Tesis para optar el grado de Magíster en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software, Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Lima, Perú.

Hoja de metadatos complementarios

- **Código ORCID del autor:** https://orcid.org/0000-0002-2449- 8937
- **Código ORCID del asesor:** https://orcid.org/0000-0001-9262-626X
- **DNI o pasaporte del autor:** 41770053
- **Grupo de investigación:** --
- **Institución que financia la investigación:** --
- **Ubicación geográfica donde se desarrolló la investigación:** Tarapoto -San Martín - San Martín, Longitud: O76°22'10.81" Latitud: S6°29'38.9"
- **Año o rango de años que la investigación abarcó:** 2018-2019

Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Vicedecanato de Investigación y Posgrado
Unidad de Posgrado

SUSTENTACIÓN DE TESIS PARA OPTAR EL GRADO ACADÉMICO DE MAGÍSTER EN INGENIERÍA DE SISTEMAS E INFORMÁTICA CON MENCIÓN EN INGENIERÍA DE SOFTWARE

En la Ciudad Universitaria, a los veintiséis (26) días del mes de febrero del 2020, siendo las 19:10 horas, se reunieron en el Auditorio de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos, el Jurado de Tesis conformado por los siguientes docentes:

Mg. Juan Carlos Gonzales Suárez (Presidente)
Mg. Zoraida Emperatriz Mamani Rodriguez (Miembro)
Dr. Augusto Bernuy Alva (Miembro)
Dr. David Santos Mauricio Sánchez (Asesor)

Se inició la Sustentación invitando al candidato a Magíster ***Richard Enrique Injante Oré****, para que realizara la exposición oral y pública de la tesis para optar el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software, siendo la Tesis intitulada:*

"Método para Recomendar Factores de Posicionamiento Personalizados en el Motor de Búsqueda de Google"

Concluida la exposición, los miembros del Jurado de Tesis procedieron a formular sus preguntas que fueron absueltas por el graduando; acto seguido se procedió a la evaluación correspondiente, habiendo obtenido la siguiente calificación:

16 DIECISEIS BUENO.

Por tanto el Presidente del Jurado, de acuerdo al Reglamento General de Estudios de Posgrado, otorga al Bachiller ***Richard Enrique Injante Oré*** *el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software.*

Siendo las 20:29 horas, el Presidente del Jurado de Tesis da por concluido el acto académico de Sustentación de Tesis.

Mg. Juan Carlos Gonzales Suárez
(Presidente)

Mg. Zoraida Emperatriz Mamani Rodriguez
(Miembro)

Dr. Augusto Bernuy Alva
(Miembro)

Dr. David Santos Mauricio Sánchez
(Asesor)

Richard Enrique Injante Ore

Método para Recomendar Factores de Posicionamiento Personalizados en el Motor de Búsqueda de Google

"Tesis presentada a la Universidad Nacional Mayor de San Marcos, Lima, Perú, para obtener el grado de Magíster en Ingeniería de Sistemas, en la mención de Ingeniería de Software"

Orientador: Dr. David Santos Mauricio Sánchez

UNMSM – LIMA

Febrero 2020

FICHA CATALOGRÁFICA

Método para Recomendar Factores de Posicionamiento Personalizados en el Motor de Búsqueda de Google

Richard Enrique Injante Ore
Lima – Perú, 2020
Orientador: Dr. David Santos Mauricio Sánchez
Disertación: Magíster en Ingeniería de Sistemas e Informática.

Universidad Nacional Mayor de San Marcos
Escuela de Posgrado
Facultad de Ingeniería de Sistemas e Informática
Unidad de Posgrado
Páginas: 116

La presente tesis está dedicado a Dios, mi esposa Katterine, mi hijo Francesc Enrique y a toda mi familia.

.

AGRADECIMIENTOS

A mi asesor Dr. David Santos Mauricio Sánchez, por su constante apoyo, orientación, dedicación y revisiones para que este trabajo cumpla con los objetivos trazados.

A mis padres, Enrique y René, quienes me dieron educación y formación, y sobre todo su amor.

A las personas que me ayudaron para terminar esta investigación y a su invalorable apoyo.

Y sobre todo doy gracias al Gran Arquitecto del Universo.

Método para Recomendar Factores de Posicionamiento Personalizados en el Motor de Búsqueda de Google

RESUMEN

El considerable aumento de sitios web en Internet con temáticas de diversa índole ha hecho que los usuarios utilicen este medio para buscar y conseguir información. Existen diferentes tipos de buscadores como los temáticos, metabuscadores y jerárquicos, pero de todos los empleados para esta tarea, la mayoría de personas emplea al buscador jerárquico de Google como su motor de búsqueda de contenidos preferido. Teniendo esto en consideración, se vuelve fundamental para los propietarios de sitios web que sus páginas web logren alcanzar las mejores posiciones en los resultados de búsqueda con el fin poder promocionarse e incrementar su número de visitas y visibilidad en Internet. Este trabajo ofrece un método basado en modelo Knowledge Discovery in Databases(KDD) y técnicas de machine learning para recomendar factores de posicionamiento personalizados a los propietarios de sitios web con el fin de que mejoren su posicionamiento en el buscador de Google. El método consta de 6 fases las cuales son: selección de los factores de posicionamiento, selección de las palabras clave, rastreo de contenido, preparación de datos, aplicación de técnica de Machine Learning y la recomendación de los factores de posicionamiento. La propuesta se aplicó en una página web de posición 46 y que a través de los factores recomendados e implementados mejoró alcanzando la posición máxima 3, en forma gradual y en un tiempo de 2 meses y 1 semana.

Palabras Clave: Factores de posicionamiento; Motor de búsqueda; Buscador de Google; Optimización en los motores de búsqueda

Method to Recommend Personalized Ranking Factors in the Google Search Engine

ABSTRACT

The considerable increase in websites on the Internet with different types of topics has led users to use this means to search and obtain information. There are different types of search engines such as thematic, metasearch and hierarchical, but of all the employees for this task, most people use Google's hierarchical search engine as their preferred content search engine. With this in mind, it becomes essential for website owners that their web pages achieve the best positions in search results in order to promote themselves and increase their number of visits and visibility on the Internet. This work offers a method based on Knowledge Discovery in Databases (KDD) model and machine learning techniques to recommend custom positioning factors to website owners in order to improve their positioning in the Google search engine. The method consists of 6 phases which are: selection of positioning factors, selection of keywords, content tracking, data preparation, application of Machine Learning technique and the recommendation of positioning factors. The proposal was applied on a website of position 46 and that through the recommended and implemented factors improved reaching the maximum position 3, gradually and in a time of 2 months and 1 week.

Keywords: Ranking Factor; Search Engine; Google Search; Search Engine Optimization

INDICE

LISTA DE TABLAS

LISTA DE FIGURAS

CAPÍTULO 1: INTRODUCCIÓN

1.1 Situación Problemática

La aparición de la Internet y la World Wide Web hizo que sea el acontecimiento más importante del siglo XX. Según Internet Live Stats (2018) existen 1.6 mil millones sitios y 4.2 Millones de usuarios de Internet, que representan el 50% de la población y la mayoría de ellos comienzan su experiencia en línea desde los motores de búsqueda. Allen (2017) menciona que los diferentes buscadores jerárquicos como Google, Baidu, Yahoo, Bing y Ask se han convertido en la forma más común entre los usuarios de acceder a la información necesaria en Internet, estos están desempeñando un papel esencial y proporciona una gran ayuda en la presentación de la información requerida de una manera más fácil, por ello los motores de búsqueda juegan un papel importante en Internet para recuperar y rankear los documentos relevantes entre una gran cantidad de páginas indexadas. Según la Figura 1 el buscador de Google se lleva la mayor parte de uso por los usuarios, el cual lo convierte en la herramienta de búsqueda más importante del mundo.

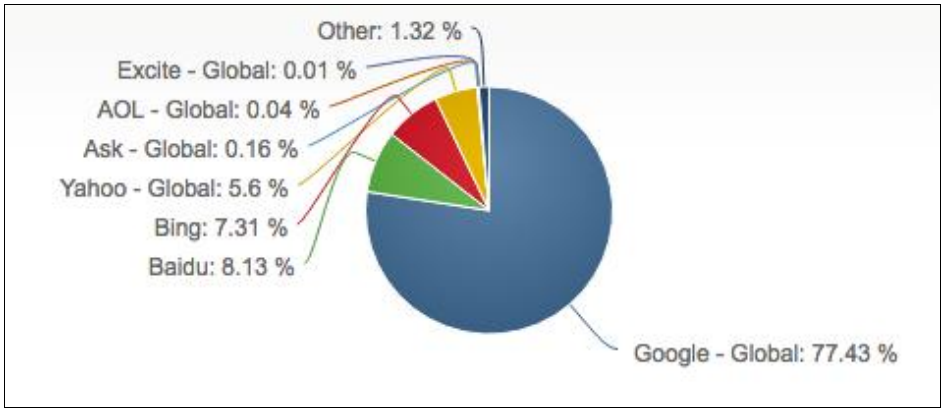


Figura 1. Cuota mundial de mercado de motores de búsqueda mundo.
Fuente. Internet Live Stats (2018).

Con el crecimiento continuo de Internet y la cantidad de los sitios disponibles, es cada vez más difícil para los sitios obtener visibilidad pública (Grzywaczewski, 2010). Prácticamente todos los propietarios de sitios web quieren tener el mayor número posible de visitantes, deseando que sus sitios

aparezcan en los primeros lugares de la página de resultados de búsqueda. Sin embargo Silva & Aguiar (2014) indican que los propietarios de sitios web requiren de conocimiento adicional, mucho esfuerzo y dedicación para lograr alcanzar visibilidad de sus sitios en el motor de búsqueda.

El principal problema para todos los propietarios de sitios web con respecto a estos motores de búsqueda es el bajo posicionamiento, el bajo tráfico y la falta de visibilidad en ellos, así que para mejorar su posicionamiento tienen que utilizar diferentes técnicas de posicionamiento, de este modo, incurre muchas veces en malas prácticas o en el uso inadecuado de estas técnicas, lo que genera entonces penalizaciones del sitio web y perjudica gravemente su posicionamiento. Muchos estudios aportan diferentes factores de posicionamiento a utilizar; sin embargo, debido al cambio frecuente del algoritmo de ranking de Google, los propietarios de sitios web corren el riesgo de aplicar factores que podrían ya estar prohibidos o desfasados, por tanto, requieren de un método que les permita identificar los factores más convenientes para su página web, esto con el fin de que utilicen estrategias adecuadas y alcancen mejores posiciones en Google. Por tal motivo, surge la necesidad de desarrollar un método que permita a los propietarios utilizar los factores de posicionamiento adecuados y, sobre todo, personalizados para su página web.

En general, el método propuesto consiste en hacer un rastreo del contenido de los resultados de búsqueda de Google bajo determinadas palabras clave y a las páginas web indexadas. Esta información servirá para obtener las métricas de los factores de las páginas web, para luego ser procesados mediante una técnica de aprendizaje de máquina y obtener las reglas que posicionen a una página web. Para finalizar, se comparan las reglas con los valores de los factores de la página web que se desee posicionar para que de esta manera se recomiende la regla más similar y los factores de posicionamiento que se debe cambiar en la página web.

1.2 Formulación del problema

¿Qué método permitirá recomendar factores de posicionamiento personalizados para que una página web pueda mejorar su posición en los resultados de búsqueda de Google?

1.3 Justificación Teórica

El aumento de los sitios web en la Internet de diferentes temáticas ha llevado al uso de los buscadores por parte de los usuarios para encontrar información y es a partir de este instante cuando es necesario aplicar técnicas de posicionamiento para que las páginas web obtengan visibilidad en el buscador. El SEO juega un papel importante como herramienta de marketing cuando se quiere aumentar la visibilidad de una página web en Internet.

Muchos estudios realizados en este ámbito han demostrado que el usuario solo visita los primeros cinco resultados antes de cambiar sus consultas. Por ejemplo, el estudio de Ochoa (2012) descubrió que “más del 80% de las primeras visitas a una página web provienen de la búsqueda web. De esas visitas, más del 76% utilizan la búsqueda de Google en todo el mundo. Además, muestra que el 84% de los usuarios de Google nunca van más allá de la segunda página de los resultados de búsqueda, y el 65% casi nunca hace clic en los resultados pagados o patrocinados”. Otro estudio como el de Allen (2017) nos dice que tan sólo los tres primeros resultados tienen una visibilidad del 79.63%, por ello es fundamental aparecer en la primera página y entre los primeros resultados orgánicos o naturales.

Existen muchos desafíos que tienen que ser considerados cuando se trabaja en el campo de SEO, en el motor de búsqueda existen varias técnicas y enfoques para mejorar la clasificación las páginas web, muchos de ellos consiste en aplicar los factores de posicionamiento propuestos por expertos y evaluar su rendimiento, pero desconocen de su influencia actual, pues debido al constante cambio del algoritmo de Google algunos factores podrían resultar negativos para la clasificación.

Bajo este contexto surge la importancia de seleccionar los factores de posicionamiento adecuados para que una página web alcance visibilidad en el motor de búsqueda.

1.4 Justificación práctica

La presente investigación busca formalizar la selección de factores de posicionamiento de manera personalizada para una página web, evitando así el abuso de técnicas de posicionamiento y obteniendo los cambios que debe realizar basado en la estructura de su sitio. Además debido al constante cambio del algoritmo de Google la propuesta permitirá mostrar los factores de posicionamiento necesarios para la página web en tiempos determinados, siendo este más efectiva que los libros o artículos sobre SEO, ya que estos solo tienen información captada en su momento del estudio. De esta manera los editores de sitios web podrán aplicar los factores con una mayor confianza y no ser penalizados por aplicar factores desfasados o desactualizados.

1.5 Objetivos

1.5.1 Objetivo General

Elaborar un método para recomendar los factores de posicionamiento personalizados para que una página web pueda alcanzar mejores posiciones en los resultados de búsqueda de Google.

1.5.2 Objetivos Específicos

- Identificar, mediante una revisión literaria, cuáles son los factores de posicionamiento internos que utilizaría Google para posicionar una página web indexada.

- Aplicar una técnica de aprendizaje supervisado para obtener las reglas que posicionen a una página web.

- Desarrollar un software que automatice el método propuesto

- Posicionar una página web aplicando el método propuesto

1.6 Propuesta

Se propone un método para la selección de los factores de posicionamiento en forma personalizada para que una página web aumente su visibilidad en el buscador de Google. Dicho método realizará un Crawling a los resultados de búsqueda de Google, de una serie de palabras clave, para extraer las páginas indexadas por orden de clasificación, luego realizará un Crawling a cada página web para extraer las métricas de los factores de posicionamiento internos, posteriormente se aplicará el algoritmo de clasificación para obtener las reglas que posicionan una página web y finalmente se compararan con las métricas de la página web a posicionar para obtener la regla más similar que recomiende los factores óptimos a cambiar para lograr aumentar su ranking. La figura 2 indica el procedimiento del método propuesto:

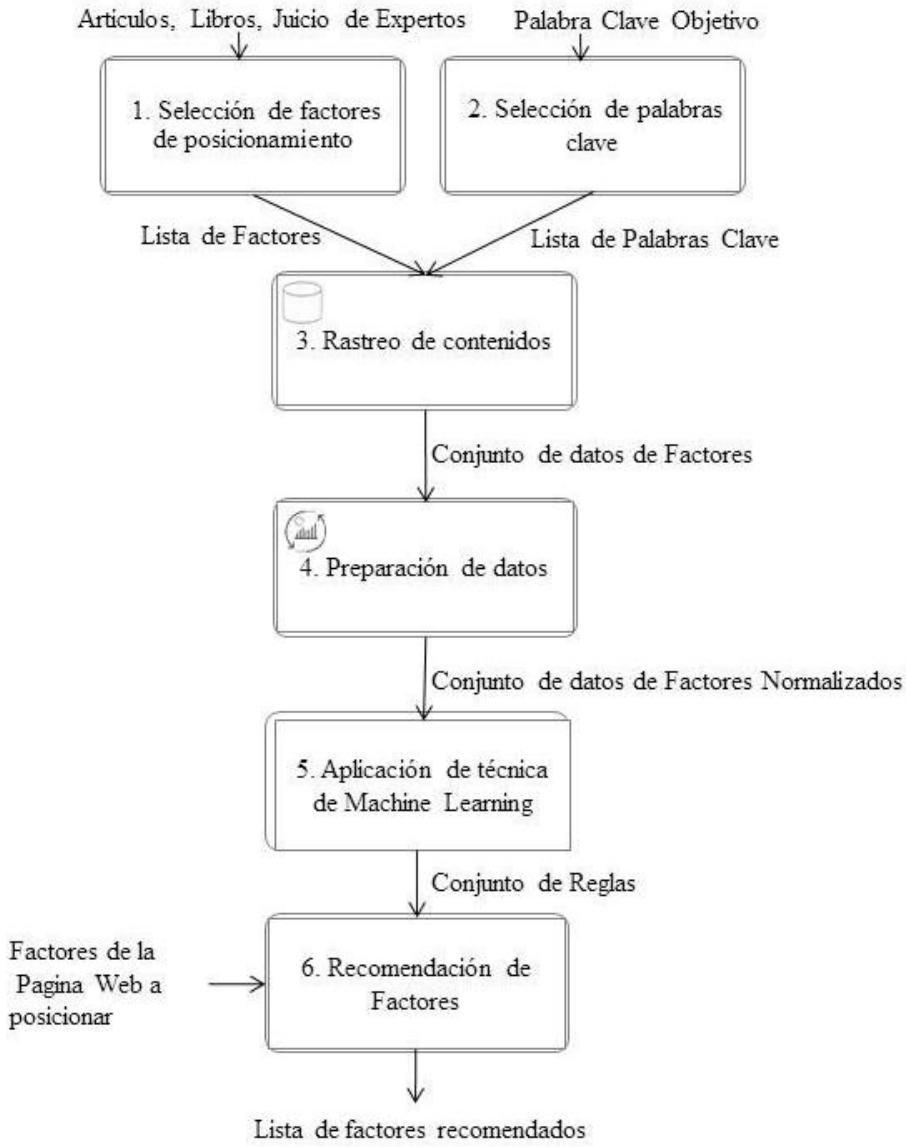


Figura 2. Procedimiento del método propuesto
Fuente. Elaboración Propia

Según la revisión literaria, muchos estudios revelan los factores que pueden influir en el posicionamiento de páginas web, pero no se ha encontrado un estudio que brinde un método que ayude a identificarlos de forma personalizada y que pueda ejecutarse en diferentes períodos con el fin de encontrar nuevos cambios que ayuden a las páginas web a posicionarse. Este método propuesto es empírico y tiene como base el proceso de

extracción de conocimiento llamado Knowledge Discovery in Databases (KDD) y está dividido en 6 fases

1.7 Organización de la Tesis

Esta tesis se encuentra organizada en siete (7) capítulos.

En el Capítulo 1 se hace la introducción la investigación, indicando el problema, la importancia, la motivación, el objetivo, la justificación y propuesta.

En el Capítulo 2 se describen las bases teóricas en donde se tocan los conceptos sobre el motor de búsqueda de Google.

En el Capítulo 3 se elabora el estado del arte de los métodos, modelos y herramientas existentes sobre optimización en los motores de búsqueda

En el Capítulo 4 se describe el aporte de esta investigación.

En el Capítulo 5 se presenta un software que automatiza el aporte

En el Capítulo 6 se muestra la validación y los resultados obtenidos del método planteado

En el Capítulo 7 se describen las conclusiones y trabajos futuros

CAPÍTULO 2: MARCO TEÓRICO

2.1 Búsqueda Web

La Búsqueda web es el acto de buscar sitios web utilizando los motores de búsqueda. Cuando alguien realiza una búsqueda utilizando los motores de búsqueda obtendrá una lista de hiperenlaces a páginas web, esta lista puede tener cien o más enlaces y el usuario puede decidir si la búsqueda contiene lo que él está buscando. Según Craswell & Hawking (2009) las páginas web son documentos web que pueden localizarse mediante un identificador denominado URL (localizador de recursos unificado), por ejemplo: **http://www.concytec.gob.pe/**. Las páginas web suelen agruparse en sitios web, conjuntos de páginas publicadas conjuntamente por ejemplo: **http://www.unmsm.edu.pe/**. La colección entera de todas las páginas web interconectadas localizadas alrededor del planeta se llama la Web, también conocida como la World Wide Web (WWW). Según Internet Live Stats (2018) Google en promedio procesa alrededor de 75.000 consultas por segundo en todo el mundo.

2.2 Motor de Búsqueda

El término “motor de búsqueda” se refiere al sitio web donde los visitantes lo utilizan para buscar documentos en Internet, pero también como un sistema utilizado para hacer “spider” en Internet, almacenar e indexar documentos web y realizar búsquedas. Los motores de búsqueda modernos son capaces de indexar una variedad de tipos de contenido, incluyendo Adobe Acrobat (pdf), Microsoft Word (doc), Microsoft Power Point (ppt), Microsoft Excel (xls) y más (Search Console Help, 2017); sin embargo, esta investigación cubrirá la indexación y recuperación de documentos web HTML.

“Los motores de búsqueda son los sistemas de información de documentos más importantes de nuestro tiempo. Constituyen una parte fundamental del panorama de Internet, aunque en los últimos años las redes sociales como Facebook y Twitter han erosionado algo de su importancia. En cualquier caso, hoy nadie podría imaginar explotar la inmensa riqueza de la Web sin la

ayuda de un motor de búsqueda, cuyas funciones no han dejado de crecer o sufrir cambios desde su aparición en los años noventa" (Pérez-Montoro & Codina, 2017).

"Los motores de búsqueda son claramente fundamentales para SEO, pero muchas organizaciones carecen de conocimiento sobre cómo funcionan. Los sitios web alojan una gama de documentos HTML, cada uno con un único URL (Uniform Resource Locator). Un motor de búsqueda permite la búsqueda en la Web creando un índice, un proceso transparente para el usuario y respondiendo a las consultas, un proceso que requiere la participación activa del usuario" (Gudivada, Rao, & Paris, 2015). Existen dos procesos (front-end y back-end) en los motores de búsqueda, el primero es la indexación y el segundo la recuperación del documento basada en la consulta del usuario.

2.2.1 Indexación

"La indexación implica buscar documentos HTML, almacenarlos en su forma original, transformar los documentos mediante procesos tales como eliminación y eliminación de palabras y generar índices y almacenarlos en una base de datos". Segun Gudivada et al (2015)

2.2.2 Consulta y recuperación de documentos

"A partir de la consulta del usuario, un algoritmo genera una lista clasificada de documentos relevantes, desde la cual el usuario busca los documentos recuperados haciendo clic en los enlaces correspondientes. Las consultas se pueden refinar y volver a ejecutar sobre la base de comentarios de los usuarios, y el motor de búsqueda almacena la información meta sobre la búsqueda actual para mejorar su rendimiento futuro. La Figura 3 muestra un esquema conceptual del proceso de consulta y recuperación de documentos". (Gudivada, Rao, & Paris, 2015).

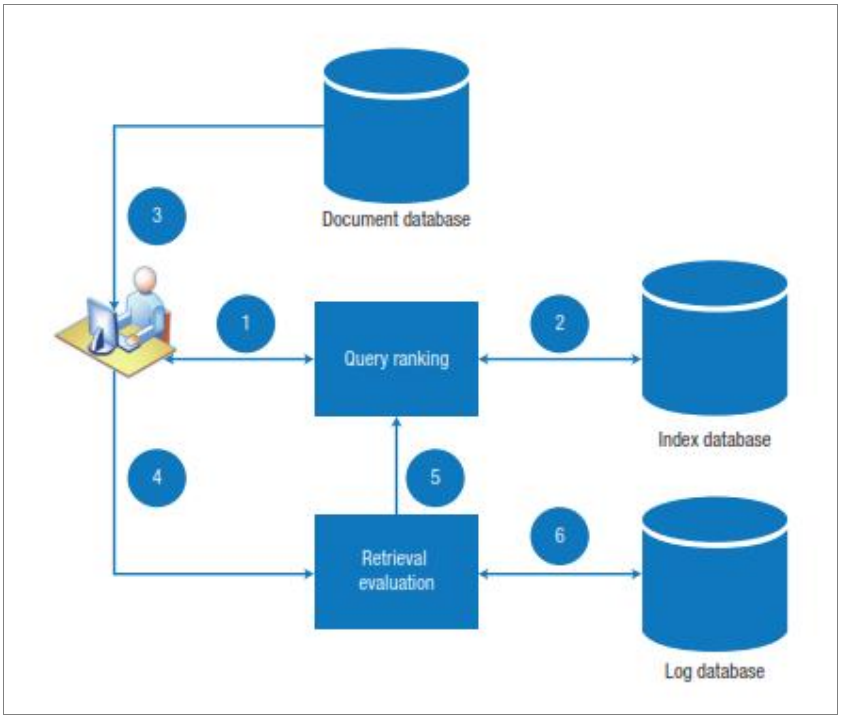


Figura 3. Proceso de consultas y recuperación de documentos.
Fuente. Gudivada, Rao, & Paris (2015)

1. El usuario emplea el navegador de un motor de búsqueda para introducir una consulta de búsqueda, normalmente una sola palabra clave o frase corta. Como en el paso 3 del proceso de indexación, el motor de búsqueda transforma la consulta del usuario en una representación canónica.
2. El algoritmo de clasificación de consultas del motor de búsqueda genera una lista clasificada de URL para los documentos que considera relevantes sobre la base de la base de datos de índices y la información contextual en la consulta del usuario. El motor de búsqueda muestra los fragmentos que corresponden a las URL clasificadas para el usuario en SERPs.
3. El usuario navega por los fragmentos y hace clic en algunos para recuperar los documentos completos correspondientes en su forma original desde la base de datos de documentos.
4. El componente de recuperación y evaluación del motor de búsqueda ayuda al usuario a refinar aún más la búsqueda sobre la base de la retroalimentación sobre la relevancia del documento: el usuario indica

explícitamente pertinencia (retroalimentación directa) o hace clic en enlaces relevantes (retroalimentación indirecta).

5. Utilizando la retroalimentación de relevancia, el motor de búsqueda puede reformular la consulta del usuario y volver a ejecutarla. Este proceso se repite hasta que el usuario esté satisfecho con los resultados de búsqueda o termine la sesión de consulta.
6. El motor de búsqueda almacena la información de metadatos, como las historias de usuario, la retroalimentación de relevancia y los fragmentos seleccionados en la base de datos de registro, que utiliza para mejorar su rendimiento de búsqueda.

"Si los resultados de búsqueda no se muestran según el interés del usuario, el motor de búsqueda se pierde su uso. Así que los algoritmos de clasificación se vuelven muy importantes". (Prabha, Duraiswamy, & Indhumathi, 2014). La Figura 4 muestra la arquitectura general del motor de búsqueda.

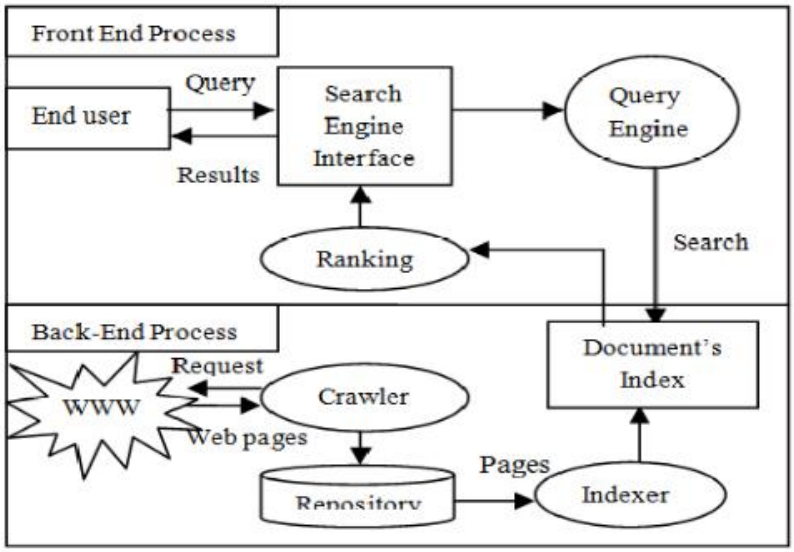


Figura 4. Arquitectura del motor de búsqueda

Fuente. Prabha, Duraiswamy, & Indhumathi (2014)

2.3 El Motor de Búsqueda de Google

“Google fue creado en 1997 como parte de un proyecto de investigación en la Universidad de Stanford, Google fue un prototipo de un motor de búsqueda a gran escala que hace un uso intenso de la estructura en el hipertexto. Google fue diseñado para rastrear e indexar la Web de manera eficiente y producir resultados de búsqueda mucho más satisfactorios que los sistemas existentes” (Brin & Page, 2012).

Google se ha convertido en la marca en la búsqueda de Internet y contiene en su índice más de 100.000.000 gigabytes. Su algoritmo PageRank da reputación a los sitios web para ofrecer los resultados más relevantes de cualquier motor de búsqueda. Google efectúa la clasificación de sus resultados de búsqueda por cada palabra clave consultada por el usuario, estos resultados contienen una lista de todos los sitios web relevantes. Google utiliza un complejo algoritmo que mantiene en secreto para mostrar lograr la clasificación final.

“La popularidad de este motor de búsqueda ha hecho el término google un verbo sinónimo de llevar a cabo una búsqueda en Internet. A pesar de la inversión pesada y la investigación de los competidores, Google sigue siendo el motor de búsqueda más popular” (Allen, 2017). Debido a su enorme éxito y popularidad, los resultados de búsqueda de los datos que se analizaran en esta investigación provienen del motor de búsqueda de Google.

2.3.1 Funcionamiento de Google

Según Carreras Lario (2014) el buscador de Google realiza su proceso de la forma siguiente**:**

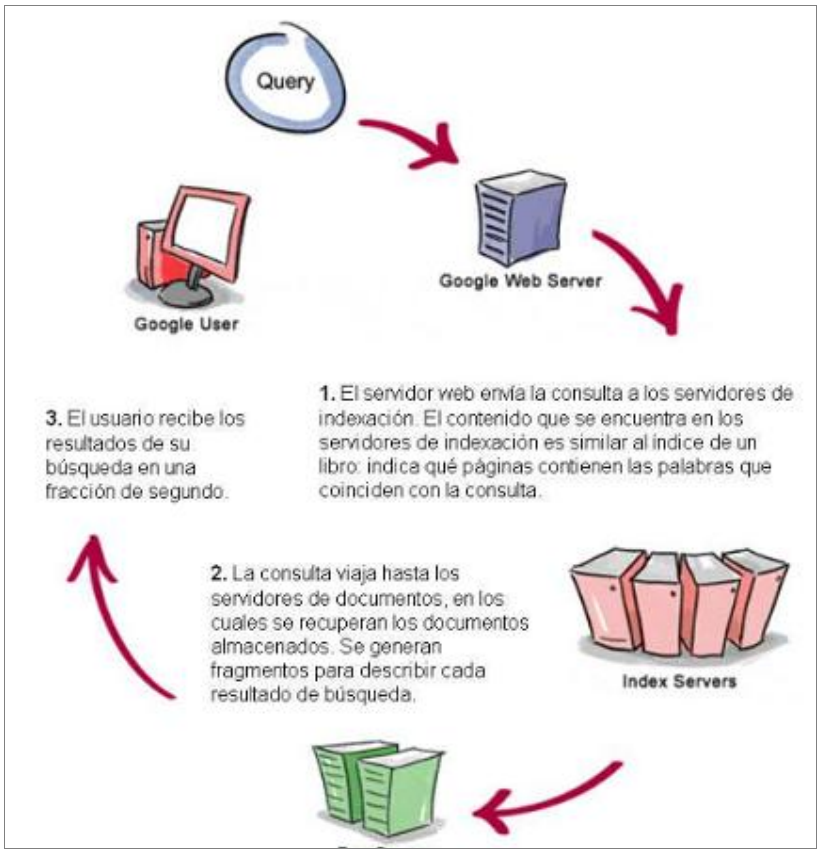


Figura 5. Proceso de Búsqueda en Google
Fuente. Carreras Lario (2014)

Al realizar una búsqueda o consulta Google genera páginas de resultados a través de su algoritmo matemático que toma en cuenta una serie de factores internos como externos.

“Cuando un usuario hace una consulta al buscador mediante una palabra clave, el algoritmo de Google examina la información en su base de datos y retorna una lista de URL que dirigen a sitios web que coinciden con lo buscado”. Ledford (2008).

Google posiciona a las páginas web en función de los puntos de PageRank que tienen. Como indican Langville & Meyer (2006), esta forma de posicionar es independiente de la búsqueda y se obtiene a través de un complejo análisis de todos las páginas web indexadas.

2.3.2 Término de Búsqueda

"Muchas personas en todo el mundo realizan búsquedas cada día mediante la presentación de términos de búsqueda en los buscadores más populares y las redes sociales. Un término de búsqueda también se conoce como palabra clave, que es la consulta textual enviada a los motores de búsqueda por los usuarios" (Palabra clave - Wikipedia, 2017).

2.3.3 Página de resultados del motor de búsqueda (SERP)

También se le conoce como SERP (Search Engine Results Page). "Una página de resultados del motor de búsqueda es la lista de resultados devueltos por un motor de búsqueda en respuesta a una consulta de palabra clave" (Página de resultados del buscador - Wikipedia, 2017). Los resultados normalmente incluyen una lista de elementos con títulos, una referencia a la versión completa y una breve descripción que muestra dónde las palabras clave han coincidido con el contenido de la página. Si vemos en SERP de Google, los elementos o listados incluidos en un SERP están creciendo en número y tipo. Algunos de los elementos de un SERP son:

- **Resultados orgánicos**: Los listados orgánicos de SERP son resultados naturales generados por los motores de búsqueda después de medir muchos factores y calcular su relevancia en relación con el término de búsqueda que desencadena. En Google, los resultados de búsqueda orgánica son páginas web de una página web que se muestran en los listados de búsqueda orgánica gratuitos de Google (Cómo medir los resultados de la búsqueda orgánica y de pago, 2017). Como se mencionó anteriormente, sólo los resultados de búsqueda orgánica se ven afectados por la optimización del motor de búsqueda, no se pagan o se "patrocinan" resultados como Google AdWords (Google, 2010)
- **Resultados pagados**: También se conocen como "patrocinados" los resultados de búsqueda que se listan en el SERP, estos se muestran en los motores de búsqueda para los clientes que pagan (propietarios

de sitios web) para aparecer en el buscador por término de búsqueda (por ejemplo, Google Adwords)

- **Gráfico de Conocimiento (Knowledge Graph)**: El Gráfico de Conocimiento es el elemento de SERP relativamente más reciente que se observa en los motores de búsqueda, particularmente en Google, para mostrar un bloque de información sobre un sujeto (El Gráfico de Conocimiento, 2017). Este listado también muestra una respuesta para preguntas de hechos tales como “Cumpleaños de Mario Vargas Llosa” o “Da Vince”.
- **Búsquedas relacionadas**: Esta parte de la SERP es donde los motores de búsqueda proporcionan sugerencia en los términos de búsqueda relacionados.

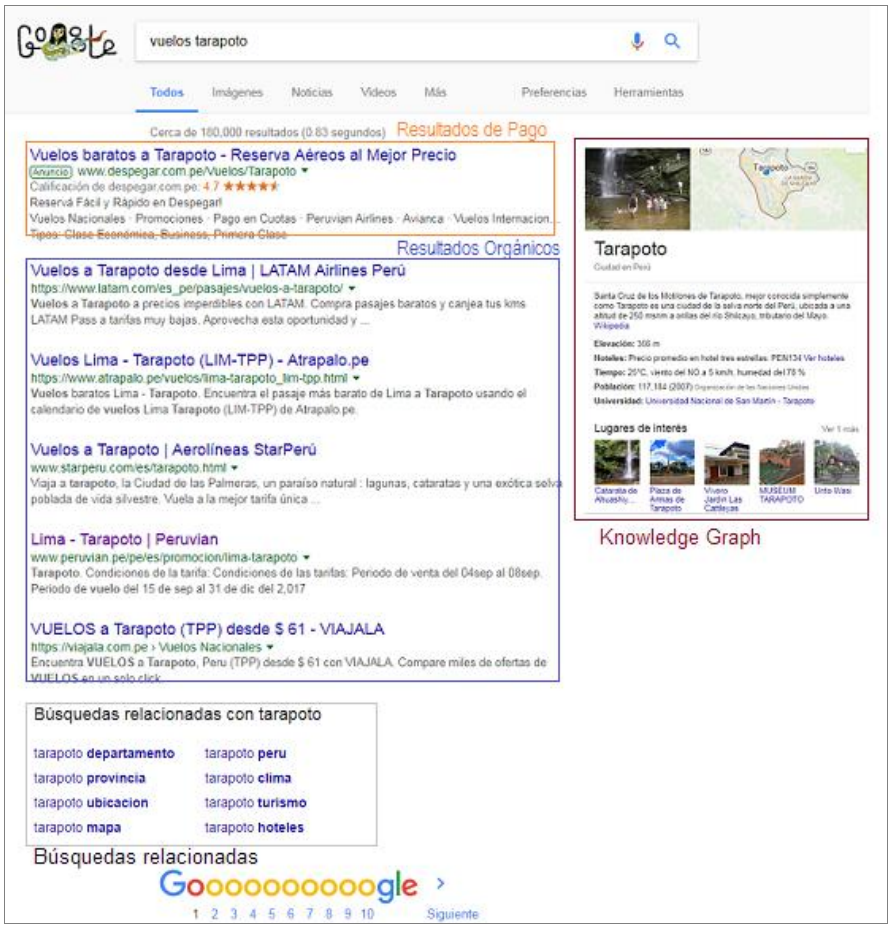


Figura 6. Página de resultados de búsqueda de Google

Fuente. Google.com.pe

2.3.4 Algoritmos del Buscador de Google

Google actualiza sus algoritmos de búsqueda varias veces al año, algunos de ellos son considerados como importantes, mientras que algunos de ellos son pequeñas actualizaciones (DiSilvestro, 2013). Es importante estar al día con las actualizaciones para hacer los cambios necesarios en las prácticas de SEO para la organización y para evitar caer en los filtros de spam de Google.

El primer Algoritmo llamado “Florida” fue lanzado por Google en noviembre de 2003 con el objetivo de detener spammers. La actualización está dirigida a la palabra clave de relleno y lo hacen completamente inútil. La larga historia de actualizaciones de algoritmos de Google ha comenzado desde entonces. Frecuentes actualizaciones de los algoritmos hacen que sea difícil para todos para el ranking web (Clarke, 2017). Uno de los principales algoritmos de búsqueda es Google Panda que fue lanzado en febrero de 2011, este algoritmo comprueba la calidad del contenido y, especialmente, se dirige al contenido de mala calidad. Evita que el contenido de calidad pobre se sitúe alto en las páginas de resultados de búsqueda de Google (SERP). Siempre que Google haga la actualización principal en este algoritmo, el sitio podría bajar tráfico orgánico, ranking o viceversa. El algoritmo Panda ha sido actualizado 28 veces desde entonces. Cambió la forma en que los profesionales crean la estrategia de contenido, la investigación de palabras clave y la estrategia de construcción de vínculos. Los enlaces entrantes relevantes de alta calidad añaden valor SEO al sitio web. (Patel, 2014). En cuanto al algoritmo de Panda sugirió que el contenido de baja calidad en la página web puede afectar la clasificación general de la página web. Por lo tanto, la fusión o la mejora del contenido de las páginas para mejorar la calidad podrían ayudar a la clasificación de los contenidos de mayor calidad. Otra actualización importante en el motor de búsqueda fue el algoritmo denominado pingüino en 2012 (Illyes, 2016), que castigó a los sitios basados principalmente en el spam de los resultados de búsqueda mediante la compra de enlaces o la obtención de enlaces sólo para

aumentar el ranking de Google. Fue actualizado por última vez el 23 de septiembre de 2016 desde el lanzamiento. Garry Illyes del equipo de ranking de búsqueda de google ha mencionado dos cambios importantes en el blog con el título “Penguin ahora es parte de nuestro algoritmo básico” (Illyes, 2016). Dos grandes cambios reconocidos en el blog fueron.

- Real-time “Los datos de Penguin se actualizan en tiempo real, por lo que los cambios serán visibles mucho más rápido, típicamente tomando efecto poco después de que rastreemos y reindexemos una página” (Illyes, 2016).
- El Penguin es ahora más granular, “Penguin ahora devalúa el spam ajustando el ranking basado en señales de spam, en lugar de afectar la clasificación de todo el sitio”. Como parte del control del spam web de Google, se realizó una actualización de las páginas en marzo de 2015 (White, 2015). Definió las páginas de entrada como “sitios o páginas creadas para clasificar altamente a consultas de búsqueda específicas. Son malos para los usuarios porque pueden llevar a varias páginas similares en los resultados de búsqueda de usuarios, donde cada resultado termina llevando al usuario esencialmente al mismo destino. También pueden llevar a los usuarios a páginas intermedias que no son tan útiles como el destino final “ (Google Search Console Help, n.d.).

El 26/10/15, Google anunció RankBrain, un sistema de inteligencia artificial de aprendizaje automático que les ayuda a ofrecer mejores resultados de búsqueda.

RankBrain es un factor de clasificación de Google, así como una máquina de aprendizaje artificial que les permite descifrar los mejores resultados de búsqueda. RankBrain es un componente del algoritmo de Google, no del algoritmo completo. Mediante el uso de inteligencia artificial, esta tecnología puede agregar información sobre la intención y la satisfacción del usuario al buscar en Google y utilizar esa información para producir mejores resultados de búsqueda. RankBrain busca contenido que coincida con las consultas de búsqueda, a pesar de que el contenido puede no tener las palabras clave exactas. Ahora, el problema es que Google ya tenía una forma de hacer coincidir una búsqueda para decir algo como: “Mejores bateadores de

béisbol en San Diego", con algo como, "Una lista de los mejores bateadores en el área".

Hasta la redacción de esta tesis la última actualización encontrada fue cuando Google sacó un nuevo algoritmo llamado Búho (Tejada, 2017) cuyo objetivo es mejorar la calidad de las sugerencias de búsqueda y detectar los resultados no deseados.

2.4 Optimización del motor de búsqueda

Entre las más relevantes definiciones podemos mencionar que SEO es:

- "Es a menudo hacer pequeñas modificaciones en su página web, como el contenido y el código. Cuando se visualizan individualmente, estos cambios pueden parecer mejoras incrementales, pero podrían tener un impacto notable en la experiencia de usuario y el rendimiento de su sitio en los resultados de búsqueda orgánica" (Google , 2010).
- "Es el proceso de mejorar la visibilidad de una página web en la página de resultados del motor de búsqueda (SERP) en respuesta a una consulta de palabras clave. La técnica de optimización de motores de búsqueda se utiliza básicamente para mejorar los resultados de búsqueda orgánica" (Duklan, Mourya, & Bahuguna, 2015)
- "Es un proceso técnico, analítico y creativo para mejorar la visibilidad de una página web en los motores de búsqueda. Su función principal es impulsar más visitas a un sitio que convertir en ventas" (Anderson, 2017)
- Según Wikipedia, "Search engine optimization (SEO) es un conjunto de métodos destinados a mejorar el ranking de una página web en los listados de motores de búsqueda"

Hay muchas definiciones de SEO, pero SEO orgánico hasta el 2018 es sobre todo conseguir tráfico libre de Google. El arte de hacer SEO radica en la comprensión de cómo la gente busca cosas y entender qué tipo de resultados Google quiere mostrar a sus usuarios. Se trata de reunir muchas cosas para buscar oportunidades. Un buen optimizador tiene una comprensión de cómo los motores de búsqueda como Google generan sus

resultados naturales para satisfacer las consultas de navegación, informativas y transaccionales de los usuarios.

Las técnicas de SEO implican dos procesos importantes: la optimización dentro de la página y la optimización fuera de la página (Rehman & Khan, 2013). La optimización dentro de la página se aplica cuando el optimizador tiene control directo en el contenido de la página web y las de fuera de la página realizan estrategias para comercializar el sitio web.

2.4.1 Importancia del SEO

Los motores de búsqueda toman casi el 90% de todo el tráfico web y al día se ejecutan millones de consultas (Internet Live Stats, 2018). Así que si la página web aparece después de la primera página en los motores de búsqueda, entonces puede perder un gran número de visitantes, porque muchos visitantes no van para la segunda página. Así que si desea aumentar el ranking de su página web en la mayoría de los populares motores de búsqueda que tiene que utilizar una combinación de palabras clave correctas, enlaces entrantes, contenido eficaz y por último, pero no menos importante, las etiquetas META. Los tres beneficios principales del SEO son los siguientes: Primero, si se gasta normalmente dinero en la publicidad para para generar tráfico al sitio web, con estrategias de SEO bien aplicadas se reducen mucho de los gastos de publicidad. En segundo lugar, encontrará que el contenido de su página web será más alto en los índices de los motores de búsqueda, lo que significa más referencias cada mes. En tercer lugar, el ranking de su sitio como un conjunto aumentará en el índice del motor de búsqueda. El tipo y la calidad del contenido determinan en gran medida el número de referencias recibidas cada mes. Con un buen SEO los mejores resultados están garantizados. Por último, SEO hará que el sitio sea más popular y otros propietarios de sitios web enlacen a su contenido. La mayor calidad y singularidad del contenido, más las referencias de enlace directo que recibirá.

2.4.2 Optimización On-page

“Optimización en la página” (a veces llamada “optimización en sitio” o ”Factores On-Page”), se produce durante el desarrollo de un sitio web. Las páginas web con las que fueron dirigidas por las palabras clave son extremadamente importantes. La investigación de palabras clave proporciona la oportunidad de comprender al público. Cada una de las páginas del sitio web debe ser únicas y optimizadas individualmente.

Para algunos temas nicho que apuntan a una frase muy específica y de bajo tráfico, los atributos “ on-page” pueden ser la única preocupación de SEO necesaria del proveedor de contenido. En muchos de estos casos, el uso cuidadoso de la palabra clave(s) en la propia página será suficiente para obtener el resultado deseado. Este es el caso ideal porque los factores “on-page” están completamente bajo el control del proveedor de contenido.

Otra estrategia empleada para ayudar a la optimización “on-page” es la vinculación interna. Los enlaces internos son hipervínculos que se dirigen al mismo dominio. Esto significa que las páginas web están internamente vinculadas entre sí. Esto permite a los usuarios navegar la web y le ayuda a distribuirse por todos enlaces (poder de clasificación) alrededor. La vinculación interna apoya a las técnicas de SEO, proporciona la experiencia del usuario y resulta en un ranking más alto.

2.4.3 Optimización Off-page

“Optimización fuera de la página” (a veces llamada ”Factores On-Page”), Conocidos como factores externos o fuera de página; son los que no pueden ser moderados por la página web para mejorar su posición (Bécares Pérez, 2013). Las estrategias fuera de página se relacionan con las prácticas en las que un sitio y sus contenidos se difunden a través de Internet para aumentar su tráfico, lo que da lugar a una mayor alta clasificación (Rehman & Khan 2013). Las influencias externas incluyen, pero no se limitan a, medios de comunicación social, publicaciones en blogs, foros, feeds RSS, comunicados de prensa, creación de vínculos, etc. Algunas de las formas sugeridas para realizar la optimización fuera de la página son: Colocar enlaces a sitios web

de redes sociales o realizar la vinculación de nuevo a sitios web de buena reputación para mejorar la clasificación de la página, así como insertar enlaces de sitios de renombre como los de “. gov “y”. edu “.

2.4.4 Técnicas SEO de sombrero blanco y de sombrero negro

Se les conoce como “White Hat SEO” y “Black Hat SEO”. Estas técnica se dividen en grupos, que van relacionadas a buenas y malas prácticas.

“Las buenas prácticas se basan en recomendaciones dadas por Google, la forma de diseñar sitios web amigables y de calidad. Mientras que las malas prácticas consisten en aplicar técnicas fraudulentas para engañar al buscador. La detección del fraude puede incurrir en la sanción por parte de Google, incluso la desindexación” (López Gómez, 2011).

Sitios web se benefician al emplear técnicas de sombrero blanco. Se supone que el enfoque de sombrero blanco sirve a los buscadores, para que los robots puedan rastrear fácilmente la información relacionada para la indexación y proporciona a los usuarios las respuestas adecuadas a sus preguntas. Estas técnicas son consideradas aceptables por los motores de búsqueda. En la siguiente tabla se exponen los pros y los contras.

Tabla 1. Pros y contras de las técnicas de sobrero blanco

Pros	**Contras**
Gratuito	Inversión a largo plazo
Confiable	Influencia desconocida
Mayor clasificación / popularidad	Falta de control
Mantener la alta clasificación por un largo tiempo	Diferentes criterios de clasificación de página entre el motor de búsqueda

Fuente. Autor

El principal beneficio de las técnicas SEO sombrero blanco es que es rentable porque no se requiere pago para colocar los sitios en los motores de búsqueda. Las estrategias que adopten las directrices de las buenas prácticas de los motores de búsqueda también se beneficiarán de ser conocidas como fiables y confiables. Involucrarse con técnicas de sombrero blanco SEO no sólo potencialmente poner el sitio en una clasificación más alta en un motor de búsqueda lista, pero también perpetuará su posición. Eventualmente, los sitios ganarán popularidad entre todos los usuarios y generarán mejores resultados para los propietarios de sitios web.

Referente a las malas prácticas o “Black-hat SEO” podemos mencionar el Cloaking que consiste mostrar un contenido diferente al robot del buscador y los usuarios. El uso de texto invisible o del mismo color de fondo. Duplicidad de dominios para copar todos los resultados del busdador y el spam de enlaces.

2.4.5 Factores de posicionamiento

Los factores de posicionamiento también conocidos como criterios de clasificación o factores de posicionamiento son los factores utilizados por buscadores para evaluar el orden de relevancia de una página web cuando alguien busca una palabra o frase en particular. Es casi obvio que los factores de posicionamiento tienen un peso diferente asignado a ellos. Por ejemplo, de acuerdo con Egri & Bayrak (2014) “El más factor importante es la duración de la estancia en el sitio, y este se encuentra en tener contenido que evite que el usuario se vaya”.

“Los motores de búsqueda funcionan mediante el uso de algoritmos para evaluar sitios web por tema y relevancia. Esta evaluación se utiliza para estructurar las páginas en el índice del motor de búsqueda, lo que en última instancia, resulta en las consultas de los usuarios que muestran la mejor clasificación posible de los resultados mostrados. Los criterios para la evaluación de páginas web y para producir esta clasificación se denominan generalmente factores de posicionamiento” (Tober, Hennig, & Furch, 2014). Las razones para esto son:

- El número cada vez mayor de documentos en Internet hace que sea imposible clasificar estas páginas sin un algoritmo automático, a pesar de la existencia de “evaluadores de calidad” humanos.
- El algoritmo es obligatorio (orden, después de todo, requiere un patrón), y, al mismo tiempo, el secreto mejor guardado en el negocio de Internet, porque para los buscadores, es primordial mantener los factores subyacentes que conforman el algoritmo estrictamente confidencial. Este secreto inherente tiene menos que ver con la competencia entre los buscadores de lo que tiene que ver con razones más básicas: si las maneras de obtener buenas clasificaciones fueran ampliamente conocidas, se volverían irrelevantes ya que serían manipuladas constantemente.

Al inicio Google consideraba las páginas relevantes para temas específicos en los que se usaban frecuentemente los términos de búsqueda asociados a temas (palabras clave). Los operadores del sitio pronto aprovecharon este conocimiento y lograron posiciones muy buenas en las SERPs al rellenar páginas con palabras clave, permitiendo que sus páginas, a menudo no relevantes, se encuentren en posiciones bien posicionadas para los términos de búsqueda buscados.

Esto generó no sólo la competencia real entre los motores de búsqueda y SEO, pero produjo el mito del factor de clasificación. El objetivo de la búsqueda semántica creó una red de criterios que inicialmente eran estrictamente técnicos (por ejemplo, el número de backlinks), pero poco después también se agregaron componentes menos técnicos (por ejemplo, señales de usuario).

Este desarrollo, junto con la búsqueda del resultado óptimo, ha culminado en la constante evolución de los factores de posicionamiento. El ciclo de retroalimentación interminable de los ciclos de actualización permanente-iterativa está diseñado exclusivamente para generar resultados de búsqueda que ofrecen mejoras constantes al buscador individual. La estructura y

complejidad de los factores de posicionamiento, sumada a la fuerte influencia de las señales de usuario, está diseñada para producir la experiencia de búsqueda más relevante para el usuario.

2.4.6 PageRank

El PageRank es algoritmo de búsqueda simple y eficiente patentado por Google (Page, 2001). Pero, debido a su simplicidad, los webmasters comenzaron a crear algoritmos y técnicas que pueden engañar el PageRank para obtener mayor clasificación en los motores de búsqueda. Debido a ello Google actualiza constantemente su algoritmo con el fin de luchar contra estas malas prácticas.

2.5 Rastreo de contenidos o Crawling

“Es una pieza de software que recorre la web de manera metódica y automatizada, es también la herramienta Base del proyecto, su implementación y configuración, permite visitar sitios web descargando la información que allí reside para luego almacenarla obteniendo resultados de búsquedas más eficientes” (Kobayashi, 2000).

El proceso de hacer Crawler inicia con la visita a las URL que se están dentro del código, estos son denominados como semillas, encuentra a los hiperenlaces de las páginas y los agrega a las URL a visitar de forma frecuente en concordancia a un grupo de reglas

```
Crawling( Conjunto de páginas iniciales S )

ColaURLs <- S
HACER {
    p <- SeleccionarURL( colaURLs )
    contenido <- Descargar( p )
    (texto, enlaces, estructura, ...)  <- Extraer( contenido )
    colaURLs <- AgregarEnlaces( colaURLs, enlaces )
} HASTA( Condición )
```

Código fuente básico del procedimiento de hacer crawler. El algoritmo muestra una serie de pasos donde se explora las URL automáticamente.

2.6 KDD como proceso para la obtención de conocimiento

"KDD (Knowledge Discovery in Databases) es una metodología genérica para encontrar información en un gran conjunto de datos y con ello generar conocimiento. Se define como un proceso no trivial de extracción de información a partir de los datos, la cual se encuentra presente de forma implícita, previamente desconocida y potencialmente útil para el usuario o para el negocio" (Fayyad et al,1996). (Velásquez & Palade, 2008).

"El objetivo principal de esta metodología es automatizar el procesamiento de los datos, permitiendo a los usuarios dedicar más tiempo a las tareas de análisis y al descubrimiento de relaciones entre los datos". Martínez Álvarez (2012).

"El KDD es un proceso que consta de una serie de etapas consecutivas, y funciona de forma iterativa e interactiva. Iterativa, ya que es posible regresar desde cualquier etapa a una anterior para ajustar los parámetros o supuestos previos, e interactiva pues el usuario experto del negocio tiene que estar presente para aportar con su conocimiento en la preparación de los datos y en la validación de los resultados que se obtengan durante el proceso." Gervilla et al (2009)

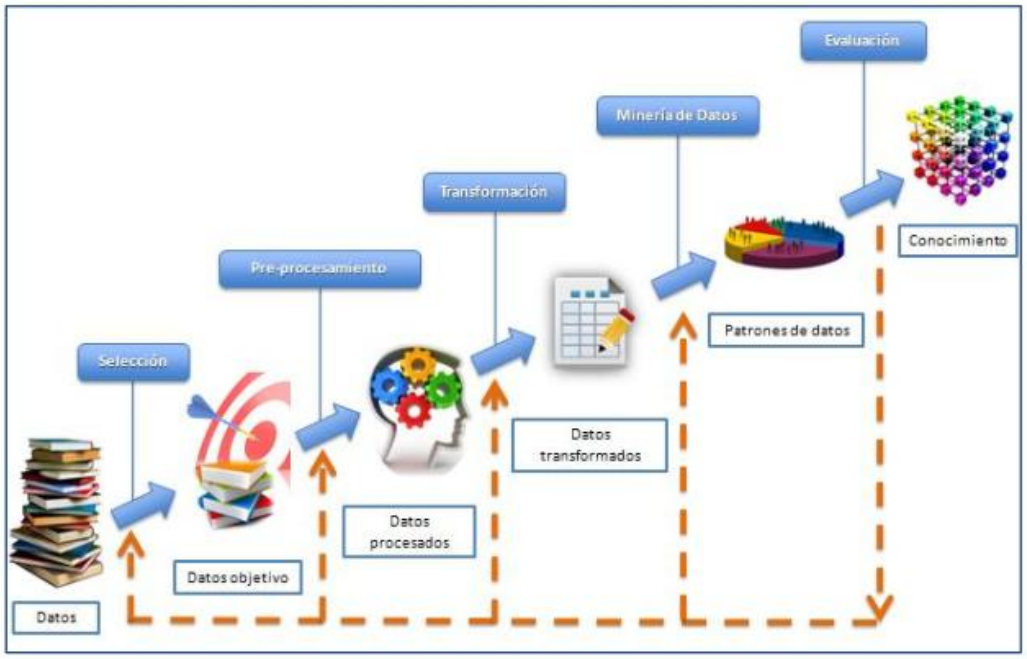


Figura 7. Proceso KDD
Fuente*. Martínez Álvarez (2012)*

"Las etapas del proceso KDD son importantes porque son la base donde se hará la minería de datos. Si los datos no fueron bien preparados, los resultados que se obtengan en el análisis no serán confiables. Por lo cual, se debe asegurar que se esté trabajando sobre una base de datos bien diseñada." Ñaupas Caraza, C. M. (2016).

2.7 Machine Learning

Machine Learning, también conocido como aprendizaje automático, sub campo de las ciencias de la computación y es rama de la inteligencia artificial, su objetivo es la desarrollar técnicas basados en algoritmos que permitan a las computadoras aprender. Trata de encontrar algoritmos para convertir los datos en programas de computadora. Los modelos o programas resultantes son capaces de generalizar comportamientos para un conjunto más amplio de datos.

Según Florián Noriega (2013) "Machine Learning puede aplicarse en tareas de Clasificación, Regresión, Ranking, Clustering y Reducción"

- Clasificación: El uso de estas técnicas es para identificar en que categoría está una entrada, por ejemplo se usan en la clasificación de diagnóstico médico, imágenes, documentos entre otros.
- Regresión: Logra la predicción de un dato real para cada un ítem, por ejemplos la predicción del clima, ventas, demanda, tasas, variables económicas entre otros.
- Ranking: usado con el fin de ordenar la información en base a criterios, como por ejemplo la lista de páginas web mostradas por un buscador.
- Clustering: estas técnicas son muy usadas en los procesos comerciales como por ejemplo para agrupar clientes o productos y ayudar a tomar decisiones a las ventas.

- Reducción: Convierte una representación inicial de baja dimensión, manteniendo sus propiedades. Por ejemplo se suele encontrar el análisis de procesamiento de imágenes.

2.8 Random Forest

Son una serie de bosques aleatorios formados por un grupo de varios árboles de clasificación. Estos están elaborados por un algoritmo que reduce la correlación entre estos árboles por dos fuentes de aleatoriedad. El algoritmo genera una predicción promediando las predicciones de cada árbol.

"Los árboles son los candidatos ideales para el bagging, dado que ellos pueden registrar estructuras de interacción compleja en los datos, con la característica de que si crecen suficientemente profundo, tienen relativamente alta imparcialidad (sin influencias de sesgos o desviaciones en la muestra)" Martínez N.(2016)

"Cada árbol se construye utilizando una muestra bootstrap aggregating (bagging) diferente de los datos originales. Alrededor de un tercio de los casos se quedan fuera de la muestra de arranque (out of bag, OOB) y no se utiliza en la construcción del árbol." Martínez N.(2016)

Ventajas:

- Maneja infinidad de variables de entrada y no excluye ninguna.
- Es el algoritmo con más certeza en la actualidad. Mientras más grande es la información mejor es su nivel de certeza.
- Velocidad de ejecución.
- Muestra estimaciones de las variables más importantes de la clasificación.
- Cuenta con un método que estima los datos perdidos y mantiene exactitud cuando grandes cantidades de datos están perdidos.
- Brinda un método que detecta como interactúan las variables.

Desventajas:

- Sus resultados de clasificación son difíciles de interpretar.
- Cuando los datos tienen atributos correlacionados con una similar relevancia los grupos más pequeños están más beneficiados que los grandes.
- Random forests se pone a favor de los atributos con más niveles para los datos que tienen variables categóricas con diferente cantidad de niveles. En consecuencia, la variable no es muy fiable para este tipo de datos.

2.9 WEKA

"WEKA se trata de un acrónimo derivado de Waikato Environment for Knowledge Analysis – Entorno para Análisis del Conocimiento de la Universidad de Waikato. Esto es porque fue esta universidad la que inició el desarrollo de Weka en 1993, no obstante, su desarrollo original fue hecho en TCL/TK y C, para en 1997 pasar a reescribirse todo el código fuente del entorno en Java, una plataforma más universal, y añadir las implementaciones de diferentes algoritmos de modelado" Calleja Gómez (2010)

Figura 8. Interfaz principal software WEKA

Fuente. WEKA

Esta herramienta cuenta con unos diferentes algoritmos para el estudio de datos, gráficas de visualización y el modelado predictivo. Su interfaz de usuario facilita el acceso a todas sus funcionalidades. También cuenta con librerías para poder integrarlas a sistemas desarrollados en JAVA. Esta herramienta para hacer minería de datos está libremente disponible bajo la GNU (licencia pública general)

CAPÍTULO 3: ESTADO DEL ARTE

Se realizará una revisión literaria referente a los diferentes métodos para identificar los factores de posicionamiento en la optimización del motor de búsqueda, para ello se seleccionarán artículos relacionados al tema para responder las cuestiones formuladas pata la investigación. Para este fin se seguirá la metodología de Kitchenham dado en 3 fases que son planeamiento, desarrollo y resultados.

3.1 Metodología de la Investigación

El procedimiento para la revisión literaria sistemática de esta tesis es el propuesto por Kitchenham (2004) y este cuenta con las siguientes etapas:

a) Planificación de la revisión: Se plantea las preguntas de la investigación y se establece el las reglas de revisión.
b) Desarrollo de la revisión: Se ejecuta el planeamiento, se buscan y seleccionan los estudios de acuerdo a los filtros de inclusión y exclusión definidos
c) Resultados de la revisión: Muestra el analisis y los resultados a las preguntas de investigación , las mismas serán presentadas en los subcapitulos 3.4 y 3.5 respectivamente

3.2 Planeamiento

Esta etapa está dirigida a responder tres preguntas de investigación a partir de la literatura existente, las cuales son:

Q1: ¿Cuáles son los factores de posicionamiento considerados en los estudios para la optimización en el motor de búsqueda de Google?

Q2: ¿Qué métodos, técnicas, modelos o algoritmos se han desarrollado para identificar los factores de posicionamiento con más relevancia?

Q3: ¿Cuáles son los factores de posicionamiento relevantes se han identificado?

Para la búsqueda literaria que se condujo a responder las preguntas planteadas se realizaron consultas a los bancos como la IEEE Xplode, ACM digital Librario, Science Direct; la investigación cubre el periodo de 2012 a 2017.

Los trabajos seleccionados responden a las siguientes cadenas de búsqueda según la tabla 2, las mismas fueron aplicadas en el título, abstract y keyword.

Tabla 2. Cadenas de búsqueda utilizadas en la Base de datos

Recurso	**Cadena de búsqueda**
ACM Digital Library	“query”: { acmdlTitle:(+Search Engine Optimization) OR recordAbstract:(+Search Engine Optimization) }, “filter”: {”publicationYear”:{ “gte”:2012 }}, {owners.owner=HOSTED} ”query”: { acmdlTitle:(+SEO) OR recordAbstract:(+SEO) }, “filter”: {”publicationYear”:{ “gte”:2012 }}, {owners.owner=HOSTED} ”query”: { acmdlTitle:(Google) AND recordAbstract:(+Google) }, ”filter”: {”publicationYear”:{ “gte”:2012 }}, {owners.owner=HOSTED}
SciELO	(ti:(“search engine optimization”)) OR (ab:(“search engine optimization”)) OR (ti:(seo)) OR (ab:(seo)) OR (ti:(google)) AND year_cluster:(“2015” OR “2016” OR “2013” OR “2010” OR “2012”) AND type:(“research-article”)
IEEE Xplode	((((((“Document Title”:Search Engine Optimization) OR “Abstract”:Search Engine Optimization) OR “Author Keywords”:Search Engine Optimization) OR “Document Title”:SEO) OR “Abstract”:SEO) OR “Author Keywords”:SEO) and refined by Year: 2012-2018
Science Direct	pub-date > 2011 and TITLE-ABSTR-KEY(Search Engine Optimization) or TITLE-ABSTR-KEY(SEO)[All Sources(Computer Science)].
Scientific.net	(titles={Web+Search+Engine+Optimization}

	&keywords={Web+Search+Engine+Optimization}&age=5 &SortBy=0 &IncludePapers=true &searchString={Web+Search+Engine+Optimization} titles={SEO} &keywords={SEO}&age=5 &SortBy=0 &IncludePapers=true &searchString={SEO})
DOAJ	{”query”:{”query_string”:{”query”:”search engine optimization”,” default_field”:”bibjson.title”, “default_operator”:”AND”}},”from”:0,”size”:10} OR {”query”:{”query_string”:{”query”:”search engine optimization”,”default_field”:”bibjson.keywords”,”default_operator” :”AND”}},”from”:0,”size”:10} OR {”query”:{”query_string”:{”query”:”search engine optimization”,”default_field”:”index.classification”,”default_operato r”:”AND”}},”from”:0,”size”:10} OR {”query”:{”query_string”:{”query”:”SEO”,”default_field”:”bibjson.titl e”,”default_operator”:”AND”}},”from”:0,”size”:10} OR {”query”:{”query_string”:{”query”:”SEO”,”default_field”:”bibjson.keyw ords”,”default_operator”:”AND”}},”from”:0,”size”:10} OR {”query”:{”query_string”:{”query”:”SEO”,”default_field”:”index.classifi cation”,”default_operator”:”AND”}},”from”:0,”size”:10}
EBSCO Discovery Service	(q=(({search engine optimization} OR {seo})) AND data>2012)

Fuente. Autor

Se consideraron los criterios de exclusión y selección establecidos en la tabla 3. Respecto a las fuentes bibliográficas de búsqueda, se han incluido artículos con factor impacto SJR.

Tabla 3. Criterios de exclusión e inclusión

Criterios de Selección	Criterios de Exclusión
• Presentan algoritmos, métodos, modelos o herramientas para la optimización en los motores de búsqueda • Propone Factores de posicionamiento • Responde a las preguntas de investigación • Proponen métricas para la medición de los factores de posicionamiento	• Papers que mencionan optimizar o crear un motor de búsqueda • Papers que trabajan con los factores off-page • Papers que definen estrategias de uso de factores de posicionamiento • Paper de revisión de literatura. • Poster, editoriales y libros.

Fuente. Autor

3.3 Desarrollo

Según la estrategia planteada los estudios identificados en la búsqueda fueron sujetos a un proceso de selección, según los criterios de exclusión e inclusión; fue de menester hacer una revisión previa del contenido para determinar la importancia con relación a la investigación y determinar si los estudios se relacionan con la selección o identificación de los factores de posicionamiento. Los resultados y el proceso se muestran en la figura 9. Posteriormente se estudiaron y analizaron los artículos obtenidos con la finalidad de dar respuesta a las cuestiones de la investigación.

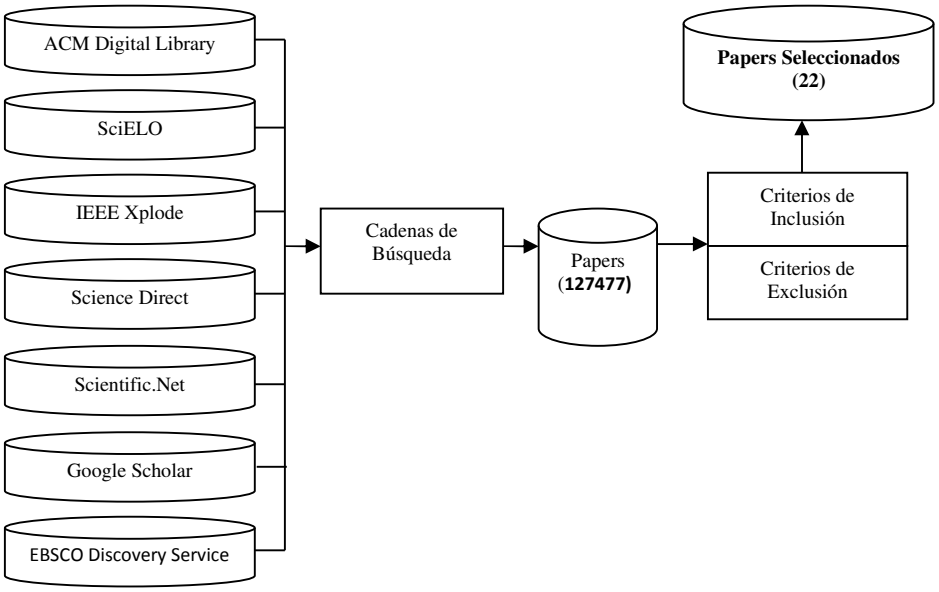


Figura 9. Proceso realizado para la revisión literaria
Fuente. Elaboración Propia

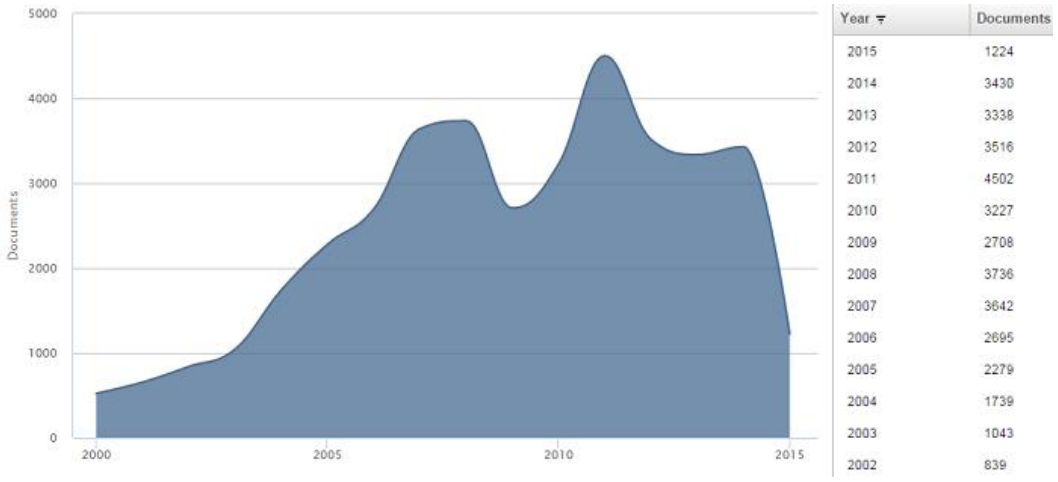
El desarrollo de toda esta revisión dio como resultado 124477 artículos, de los cuales 22 documentos fueron seleccionados por estar relacionadas con la optimización en los motores de búsqueda.

3.4 Resultados

3.4.1 Visión de las publicaciones

La figura 10 y 11 muestran el total de publicaciones en los últimos 5 años que cubren el área de motores de búsqueda y la optimización en los motores de búsqueda.

Palabra Clave: Search Engine (39096 publicaciones)



Year	Documents
2015	1224
2014	3430
2013	3338
2012	3516
2011	4502
2010	3227
2009	2708
2008	3736
2007	3642
2006	2695
2005	2279
2004	1739
2003	1043
2002	839

Figura 10. Visión temporal de publicaciones motores de búsqueda

Fuente. Scopus

Palabra Clave: Web SEO (125 publicaciones)

Year	Documents
2015	8
2014	19
2013	19
2012	18
2011	18
2010	16
2009	14
2008	5
2007	3
2006	2
2005	3

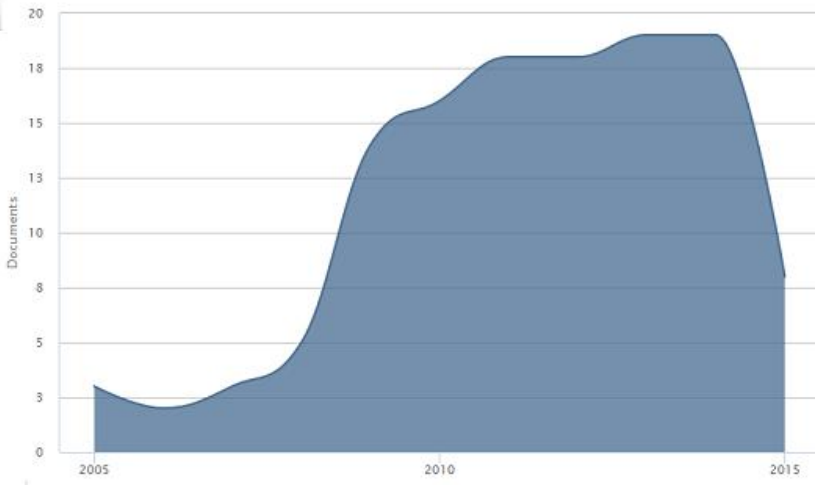
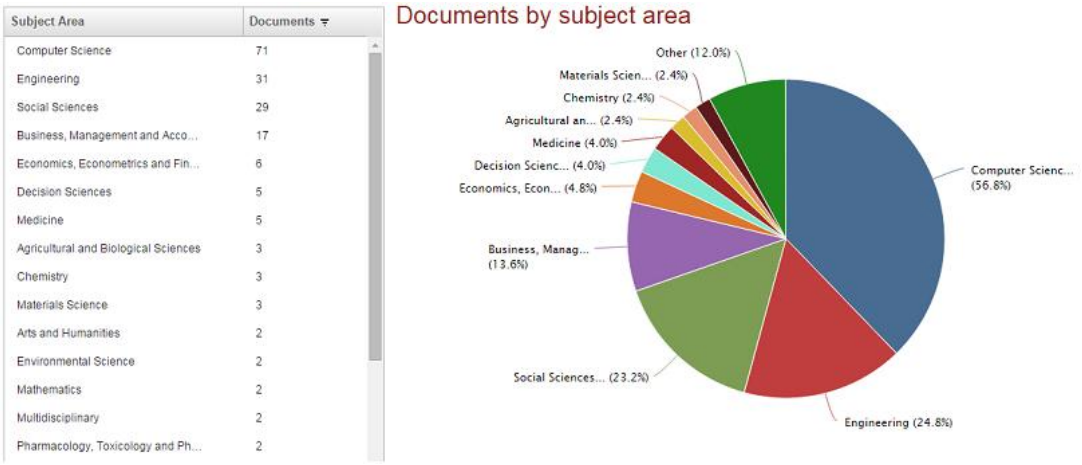


Figura 11. Publicaciones sobre SEO por año.

Fuente. scopus.com



Subject Area	Documents
Computer Science	71
Engineering	31
Social Sciences	29
Business, Management and Acco...	17
Economics, Econometrics and Fin...	6
Decision Sciences	5
Medicine	5
Agricultural and Biological Sciences	3
Chemistry	3
Materials Science	3
Arts and Humanities	2
Environmental Science	2
Mathematics	2
Multidisciplinary	2
Pharmacology, Toxicology and Ph...	2

Figura 12. Publicaciones por área de conocimiento sobre SEO

Fuente. scopus.com

3.4.2 Publicaciones por Fuentes

Los resultados muestran que la IEEE Xplode cuenta con más publicaciones relacionadas al tema de investigación como se puede apreciar en la tabla 4.

Tabla 4. Publicaciones realizadas por fuente de referencia sobre optimización en los motores de búsqueda 2012-2017

Fuente	**Estudios potencialmente elegibles**	**Estudios Seleccionados**
SciELO	75	2
ACM Digital Library	80055	1
Science Direct	33374	3
IEEE Xplode	6416	7
Scientific.Net	3236	3
Google Escolar	16400	4
EBSCO Discovery Service	80969	2
TOTAL	**127477**	**22**

Fuente. Autor

3.5 Análisis

Por lo general el proceso para optimización en los motores de búsqueda consiste en una secuencia de 5 fases (Roldán, 2017), la figura 13 ilustra este proceso.

Figura 13. Proceso de optimización en el motor de búsqueda
Fuente. Roldán (2017)

La Fase 1 tiene por finalidad descubrir al cliente conociendo sus objetivos y características del negocio así como el análisis de todo el sitio web.

La Fase 2 consiste en buscar los criterios o palabras clave más relevantes para el negocio, análisis de la competencia y ueso de herramientas generadoras de estas.

La Fase 3 consiste en asegurar que se cumpla al máximo las directrices de calidad contenido y estructura de la misma, además la de evaluar la relevancia que tiene el contenido para el usuario según su criterio de búsqueda.

La Fase 4 consiste en el uso de herramientas para medir el tráfico a la página web y saber qué lo genera con el fin de formular nuevas estrategias.

La Fase 5 consiste en medir la página web y hacerle seguimiento periódicamente con el fin de evaluar el sitio, hacer ajustes y el cumplimiento de los objetivos.

La presente investigación se centrará específicamente en la fase de Optimización de Contenido, puesto que es quien contiene los factores de posicionamiento on-page a investigar. En los medios on-line en contenido de la página web es muy importante y en esta fase se debe asegurar que se cumplan los factores de posicionamiento adecuados. En este sentido se debe asegurar que el contenido contenga los factores necesarios para que una página web mejore su posicionamiento.

Diferentes investigaciones proponen estrategias, métodos, modelos o algoritmos, algunas se orientan al contenido off-page, on-page y otros a las palabras clave, todas con el fin de mejorar el posicionamiento de las páginas web e identificar los factores de posicionamiento que influyen en estas. A continuación se revisará la literatura seleccionada con el fin de responder las preguntas de investigación.

Q1: ¿Cuáles son los factores de posicionamiento considerados en los estudios para la optimización en el motor de búsqueda de Google?

Los factores de posicionamiento también conocidos como criterios de clasificación o factores de posicionamiento son los criterios utilizados por los buscadores para evaluar la relevancia de una página web cuando alguien busca una palabra o frase en particular. Los motores de búsqueda funcionan mediante el uso de algoritmos (que son un secreto) que evalúan la página web por tema y relevancia. Esta evaluación se utiliza para estructurar las páginas en el índice del motor de búsqueda, lo que en última instancia, resulta en las consultas de los usuarios que muestran la mejor clasificación posible de los resultados mostrados. Los criterios para la evaluación de páginas web y para producir esta clasificación se denominan generalmente factores de posicionamiento (Tober, Hennig, & Furch, 2014).

Para considerar los factores de posicionamiento los autores utilizaron diferentes técnicas y herramientas como la revisión literaria, juicio de expertos y servicios de análisis de sitios web.

En la Tabla 5 se listan los diferentes factores de posicionamiento considerados en los las investigaciones seleccionadas.

Tabla 5. Lista de factores de posicionamiento considerados en los trabajos para la optimización en el motor de búsqueda

Factores	**Referencias**
Keyword en el título *Palabra clave en la etiqueta title*	(Nasomyon & Wisitpongphan, 2014), (Sylvain Sagot & Fougèr, 2016), (Krrabaj, Baxhaku , & Sadrijaj, 2017), (Gupta, Rakesh, Thakral, & Chaudhary, 2016), (Eswarawaka, Kudikala, Kuchi, & Verma K, 2017), (Shubham, Shubham Soni, & Shweta, 2015), (Moráguez & Cancio, 2014), (Al-Jadaan, 2015), (Al-Jadaan, 2015), (Hussien, 2014), (Aregay, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013), (Bécares Pérez, 2013)
Backlinks *Número de Enlaces externos al sitio web*	(Nasomyon & Wisitpongphan, 2014), (Lin & Chi , 2014), (Eswarawaka, Kudikala, Kuchi, & Verma K, 2017), (Sandhya , Thakare, & Butey, 2016), (Shubham, Shubham Soni, & Shweta, 2015), (Duklan, Mourya, & Bahuguna, 2015), (Sarika & Sharma, 2014),(Al-Jadaan, 2015), (Aregay, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Keyword in URL *Palabra clave en la URL de la página web*	(Nasomyon & Wisitpongphan, 2014), (Krrabaj, Baxhaku , & Sadrijaj, 2017), (Gupta, Rakesh, Thakral, & Chaudhary, 2016), (Shubham, Shubham Soni, & Shweta, 2015), (Silva & Aguiar,

	2014), (Moráguez & Cancio, 2014), (Al-Jadaan, 2015), (Hussien, 2014), (Aregay, 2014), (Bécares Pérez, 2013)
Keyword in Meta Keyword *Palabra clave en la meta etiqueta Keyword*	(Nasomyon & Wisitpongphan, 2014), (Duklan, Mourya, & Bahuguna, 2015), (Sarika & Sharma, 2014), (Moráguez & Cancio, 2014), (Al-Jadaan, 2015), (Hussien, 2014), (Aregay, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013), (Bécares Pérez, 2013)
External Links *Números de enlaces externos*	(Nasomyon & Wisitpongphan, 2014), (Eswarawaka, Kudikala, Kuchi, & Verma K, 2017), (Themistoklis & Symeonidis , 2015), (Moráguez & Cancio, 2014),(Al-Jadaan, 2015), (Aregay, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Google + *Si la página web tiene Google+*	(Nasomyon & Wisitpongphan, 2014), (Gupta, Rakesh, Thakral, & Chaudhary, 2016), (Shubham, Shubham Soni, & Shweta, 2015), (Themistoklis & Symeonidis , 2015), (Moráguez & Cancio, 2014),(Al-Jadaan, 2015), (Aregay, 2014)
Internal Links *Números de enlaces internos*	(Nasomyon & Wisitpongphan, 2014), (Silva & Aguiar, 2014), (Themistoklis & Symeonidis , 2015), (Moráguez & Cancio, 2014), (Al-Jadaan, 2015), (Aregay, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Keyword in H1 *Palabra clave en la etiqueta H1*	(Nasomyon & Wisitpongphan, 2014), (Sylvain Sagot & Fougèr, 2016), (Sarika & Sharma, 2014), (Al-Jadaan, 2015), (Hussien, 2014), (Aregay, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013), (Bécares Pérez, 2013)
Keyword in Meta Description *Presencia de la palabra*	(Nasomyon & Wisitpongphan, 2014), (Gupta, Rakesh, Thakral, & Chaudhary, 2016), (Duklan, Mourya, & Bahuguna, 2015), (Al-Jadaan, 2015),

clave en el meta description	(Hussien, 2014), (Aregay, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013), (Bécares Pérez, 2013)
Facebook *Si la página web tiene Fanspage*	(Nasomyon & Wisitpongphan, 2014), (Gupta, Rakesh, Thakral, & Chaudhary, 2016), (Shubham, Shubham Soni, & Shweta, 2015), (Moráguez & Cancio, 2014),(Al-Jadaan, 2015), (Aregay, 2014)
Keyword en ALT de Imágenes *Palabra clave en el atributo ALT del tag IMG*	(Sandhya , Thakare, & Butey, 2016), (Silva & Aguiar, 2014), (Moráguez & Cancio, 2014), (Al-Jadaan, 2015), (Hussien, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013), (Bécares Pérez, 2013)
Keyword in H2 *Presencia de la palabra clave en el tag H2*	(Nasomyon & Wisitpongphan, 2014), (Sarika & Sharma, 2014), (Moráguez & Cancio, 2014), (Al-Jadaan, 2015), (Aregay, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Keyword in H3 *Presencia de la palabra clave en el tag H3*	(Nasomyon & Wisitpongphan, 2014), (Sarika & Sharma, 2014), (Al-Jadaan, 2015), (Aregay, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Nro de páginas indexadas *Número de páginas indexadas a Google*	(Eswarawaka, Kudikala, Kuchi, & Verma K, 2017), (Moráguez & Cancio, 2014), (Al-Jadaan, 2015), (Aregay, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Robots.txt *Si la página web tiene el archivo robots.txt*	(Nasomyon & Wisitpongphan, 2014), (Moráguez & Cancio, 2014), (Al-Jadaan, 2015), (Hussien, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Keyword in H4 *Palabra clave en la etiqueta H4*	(Nasomyon & Wisitpongphan, 2014), (Sarika & Sharma, 2014), (Al-Jadaan, 2015), (Aregay, 2014)
Keyword in H5 *Palabra clave en la*	(Nasomyon & Wisitpongphan, 2014), (Sarika & Sharma, 2014), (Al-Jadaan, 2015), (Aregay, 2014)

etiqueta H5	
Keyword in H6 *Palabra clave en la etiqueta H6*	(Nasomyon & Wisitpongphan, 2014), (Sarika & Sharma, 2014), (Al-Jadaan, 2015), (Aregay, 2014)
PageRank *Calificación PageRank*	(Themistoklis & Symeonidis , 2015), (Moráguez & Cancio, 2014), (Al-Jadaan, 2015), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Sitemap *Si la página web contiene sitemaps.xml*	(Nasomyon & Wisitpongphan, 2014), (Moráguez & Cancio, 2014), (Al-Jadaan, 2015), (Hussien, 2014)
Velocidad de carga del sitio *Tiempo de carga de la página web*	(Chhabra, Mittal, & Sarkar, 2016), (Gupta, Rakesh, Thakral, & Chaudhary, 2016), (Sarika & Sharma, 2014), (Egri & Bayrak, 2014)
Edad del sitio *Tiempo del dominio activo*	(Nasomyon & Wisitpongphan, 2014), (Moráguez & Cancio, 2014), (Al-Jadaan, 2015)
Keyword Repetition *Veces que se repite la palabra clave*	(Nasomyon & Wisitpongphan, 2014), (Sylvain Sagot & Fougèr, 2016), (Aregay, 2014)
NoFollow Presencia del atributo NoFollow	(Moráguez & Cancio, 2014), (Al-Jadaan, 2015), (Aregay, 2014)
Responsive Web Design *Si la página web es responsivo o adaptable a dispositivos móviles*	(Krrabaj, Baxhaku , & Sadrijaj, 2017), (Gupta, Rakesh, Thakral, & Chaudhary, 2016), (Sarika & Sharma, 2014)
Twitter *Si la página web cuenta con Twitter*	(Nasomyon & Wisitpongphan, 2014), (Moráguez & Cancio, 2014),(Al-Jadaan, 2015), (Aregay, 2014)
Dmoz Directory	(Nasomyon & Wisitpongphan, 2014), (Morato,

Si la página web está registrado en el directorio Dmoz	Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Fresco *Última actualización del documento*	(Moráguez & Cancio, 2014), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Tamaño de la página *Peso en Kb del documento*	(Al-Jadaan, 2015), (Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Usar tag <H1> al <H6> *Presencia de las etiquetas <H1> al <H6>*	(Gupta, Rakesh, Thakral, & Chaudhary, 2016),(Al-Jadaan, 2015), (Bécares Pérez, 2013)
Validación W3C *Puntaje del validador de la W3C*	(Sarika & Sharma, 2014), (Moráguez & Cancio, 2014)
AlexaRank *Ranking en Alexa*	(Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Autoridad Dominio MOZ *Ponderado dado por MOZ*	(Themistoklis & Symeonidis , 2015)
Autoridad Pagiba MOZ *Ponderado dado por MOZ*	(Themistoklis & Symeonidis , 2015)
Extensión *Extensión de dominio, ejemplo .com, .net, org*	(Moráguez & Cancio, 2014)
Flow rate Alexa *Ponderado dado por Alexa*	(Lin & Chi , 2014)
Google analytics code *Presencia del código de*	(Aregay, 2014)

Google Analytics	
Guiones en URL *Presencia de guiones en la URL*	(Hussien, 2014)
HTML5 *Uso de HTML5 en la página*	(Hussien, 2014)
HTTPS *Uso del protocolo HTTPS*	(Gupta, Rakesh, Thakral, & Chaudhary, 2016)
Identificador del Idioma *Código que especifica el idioma de la página web*	(Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Keyword en el dominio *Palabra clave en el dominio*	(Al-Jadaan, 2015), (Aregay, 2014)
Keyword en el title tag <A> *Palabra clave en el atributo title del tag <A>*	(Al-Jadaan, 2015)
Keyword en links *Palabra clave en el tag <A>*	(Aregay, 2014)
Keyword en negrita *Palabra clave resaltado con negrita*	(Aregay, 2014), (Bécares Pérez, 2013)
Link of Wikipedia *Cuenta con enlace desde Wikipedia*	(Nasomyon & Wisitpongphan, 2014)
MozRank	(Themistoklis & Symeonidis , 2015)

Ponderación dado por MOZ	
MozTrust	(Themistoklis & Symeonidis , 2015)
Ponderación dado por MOZ	
Renovación dominio	(Moráguez & Cancio, 2014)
Años para el vencimiento del dominio	
Sitio en Flash	(Hussien, 2014)
Si el sitio está completamente en Flash	
Tasa de rebote	(Egri & Bayrak, 2014)
Tasa de rebote del sitio	
Tiempo de sesión	(Egri & Bayrak, 2014)
Tiempo de permanencia del usuario en el sitio web	
URL canónico	(Al-Jadaan, 2015)
La URL del dominio está canonizado	
Contenido Duplicado	(Gupta, Rakesh, Thakral, & Chaudhary, 2016)
Porcentaje de contenido duplicado	
Delicious Index	(Nasomyon & Wisitpongphan, 2014)
Si el sitio está Indexado en Delicious	

Fuente. Autor

Como podemos notar los factores más utilizados (más de 5 referencias) son los siguientes: Keyword in Title , Keyword Density Document, Backlinks, Keyword in URL, Keyword in Meta Keyword, External Links, Google + , Internal Links, Keyword in H1 , Keyword in Meta Description, Facebook,

Keyword en ALT de Imágenes, Keyword in H2 , Keyword in H3, Nro de páginas indexadas y Robots.txt

Cabe mencionar que la métrica de cada factor fue extraída de forma diferente para cada autor, por ejemplo algunos consideran solo la presencia de la etiqueta <H1> con un valor boleado, mientras que otros consideran si la palabra clave se encuentra dentro de la misma o el número de veces que aparece la palabra.

Q2: ¿Qué métodos, modelos o algoritmos se han desarrollado para identificar los factores de posicionamiento más relevantes?

Los métodos, modelos o algoritmos son los procedimientos que siguen los autores para conseguir los factores clasificación más influyentes, en la tabla 6 se muestran los diferentes trabajos con sus respectivas técnicas.

Tabla 6. Lista de métodos, modelos o algoritmos empleados por autor

Métodos	**Referencias**
Árbol de Decisiones algoritmo J48	(Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Cálculo Simple	(Hussien, 2014)
CfsSubsetEva/Bestfirst	(Morato, Sánchez-Cuadrado, Moreno & Moreiro, 2013)
Estadística Descriptiva	(Moráguez & Cancio, 2014)
Estudio Comparativo de revisión literaria	(Chhabra, Mittal, & Sarkar, 2016), (Sandhya , Thakare, & Butey, 2016), (Sarika & Sharma, 2014), (Gudivada, Rao, & Paris, 2015)
Estudio/Evaluación Empírica	(Gupta, Rakesh, Thakral, & Chaudhary, 2016), (Krrabaj, Baxhaku , & Sadrijaj, 2017),(Shubham, Shubham Soni, & Shweta, 2015), (Egri & Bayrak, 2014), (Al-Jadaan, 2015)
Fuzzy Decision Support	(Sylvain Sagot & Fougèr, 2016)

Systems	
Ingeniería Inversa	(Bécares Pérez, 2013)
K-means	(Lin & Chi , 2014), (Duklan, Mourya, & Bahuguna, 2015)
Método investigación-acción	(Silva & Aguiar, 2014)
Coeficiente de correlation de Pearson	(Nasomyon & Wisitpongphan, 2014), (Aregay, 2014)
Coeficiente de correlation Point-biserial	(Nasomyon & Wisitpongphan, 2014), (Aregay, 2014)
Six Sigma - DMAIC	(Eswarawaka, Kudikala, Kuchi, & Verma K, 2017)
Spearman correlation coefficient	(Themistoklis & Symeonidis , 2015), (Aregay, 2014)
TF-IDF	(Lin & Chi , 2014)
Experimental Pre-test, post-test	(Rayhan, 2013)
Support vector machine	(Su, Hu, Kuzmanovic, & Koh, 2014)
Linear Programming	(Su, Hu, Kuzmanovic, & Koh, 2014)

Fuente. Autor

Como se aprecia los estudios empíricos son los más utilizados en los diferentes trabajos seguidos por los estudios comparativos en los cuales revisan diferentes fuentes bibliográficas y determinan los factores más importantes.

Q3: ¿Qué factores de posicionamiento relevantes se han identificado?

Diferentes estudios se han abocado a identificar cuáles son los factores de posicionamiento que más influyen para que una página web se posicione mejor en los resultados del motor de búsqueda, entre los más importantes tenemos a Egri y Bayrak (2014) quienes utilizaron herramientas como PageSpeedInsights y Pingdom para medir el tiempo de carga, velocidad, tasa de rebote, vistas de página y diseño de la página para mantener al usuario en el sitio. Además, efectuaron un análisis con Google Analytics e

identificaron que el factor de posicionamiento más importante es el tiempo de duración del usuario en el sitio y este está influenciado directamente con la rapidez de carga de la web. Por otro lado, Lin y Chi (2014) propusieron un método que utiliza la frecuencia de término-frecuencia inversa de documento (TF-IDF) y K-means para identificar la combinación de palabras clave que beneficiarán la optimización del motor de búsqueda; como resultado, la página web de su estudio recibió un importante impacto reflejado en diversos indicadores, entre los que destaca su mejora en el ranking Alexa, y en factores, como el número de backlinks. Hussien (2014) llevó a cabo una investigación empírica mediante la ponderación de 20 factores basados en una ecuación propuesta, recomendando el uso de guiones en el localizador de recursos uniforme (URL) del sitio, minimizar los errores ortográficos, utilizar encabezados H1, una meta descripción adecuada y uso de sustantivos en la página. Moráguez, M. y Cancio (2014), a diferencia de los otros autores, emplearon una encuesta para identificar cuáles son los factores que influyen en tener un bajo posicionamiento en los sitios web de especialidades médicas, ellos encontraron que el uso inadecuado de las palabras clave en las etiquetas Meta, enlaces internos que no facilitan la navegación del usuario, poca actualización de documentos y contenidos del sitio muy diseminados afectan considerablemente el posicionamiento. Duklan y otros (2015) identificaron factores que tienen máximo impacto en el ranking. Para este propósito, usaron el análisis de clúster de k-means para agrupar los factores externos, así, se obtuvo que el intercambio de enlaces, Metatags, publicación en directorios, seguimiento de sitios web, cumplimiento de normas W3C y la presentación de marcadores sociales, mejora el posicionamiento de una página web. Krrabaj y otros (2017) estudiaron diferentes factores gracias al uso de herramientas proporcionadas por Google para evaluar su sitio. En este caso, el análisis fue dirigido a una página web educativo, donde identificaron que los factores de mayor impacto son el uso de la palabra clave en la etiqueta título del artículo y también en la URL. Además, las palabras clave deben aparecer al menos tres veces en el contenido principal. Finalmente, Eswarawaka, Kudikala y otros (2017) revisaron diferentes factores de posicionamiento y proponen un método para hacer SEO, apoyado en la metodología Sigsixma,

en la que experimentan con las palabras clave dentro de la etiqueta título y el documento, donde concluye que el contenido debe estar directamente relacionado con lo que el usuario está buscando, es decir, que la palabra clave aparezca con mayor frecuenta en el documento.

CAPÍTULO 4: PROPUESTA DEL METODO DE POSICIONAMIENTO

Debido a que en la revisión literaria no se ha encontrado un método para recomendar factores de posicionamiento web de forma personalizada, se propone uno que consiste en utilizar recuperar documentos de páginas web de las mejores posiciones y malas posiciones con el fin de aplicar la técnica de Random Forest, obtener reglas, comparar reglas y proponer a una página web que factores de posicionamiento debe considerar para posicionarse en el buscador de Google en una temática en particular.

4.1 Método

La siguiente figura muestra, en general, el procedimiento del método propuesto.

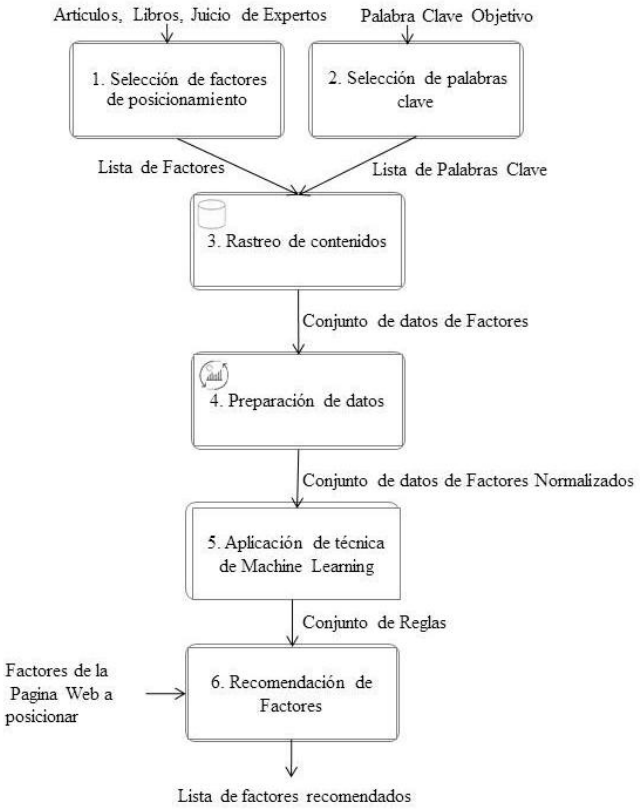


Figura 14. Procedimiento del método propuesto

Fuente. Elaboración Propia

La figura 14 nos muestra el método a seguir paso a paso para lograr identificar los factores de posicionamiento que permitan posicionar a una página web; este procedimiento inicia con la selección de palabras clave de una temática determinada para luego extraer, mediante un crawler, las URL's y métricas de los sitios web que aparecen en los resultados del motor de búsqueda, pre-procesamos los datos y aplicamos el algoritmo de clasificación Random Forest con el fin de obtener reglas que luego serán comparadas con la página web a posicionar; y así obtener las recomendaciones que mejoren su clasificación en el motor de búsqueda.

El método propuesto genera la regla más útil para que la página web mejore sus posiciones en el motor de búsqueda. En comparación con otras investigaciones, donde solo recomiendan los factores a utilizar en forma general, este método da recomendaciones personalizadas, evitando así aplicar factores innecesarios o poco relevantes para la página web, también es aplicable para sitios web relacionados a la misma temática. Otra ventaja del método es que se puede aplicar en el tiempo adaptándose así a los cambios del algoritmo del motor de búsqueda de Google.

4.2 Fases

A continuación se describen a detalle las fases del método propuesto, entrada, procedimiento y salida, además de las herramientas creadas y utilizadas para el estudio.

4.2.1 Fase 1. Selección de los Factores de Posicionamiento

Los factores de posicionamiento son aquellos que utilizan los buscadores para evaluar la relevancia ordenada de una web cuando alguien la busca basándose en una palabra clave, por lo cual es de presumir que los factores de posicionamiento tienen un peso diferente asignado y que estos son utilizados por el algoritmo de Google (Aregay, 2014). En esta fase se deben seleccionar los factores de posicionamiento a considerar para la ejecución del método, esto con el fin de generar una lista que se usará en la fase 3. La

tabla 7 muestra las entradas, herramientas y técnicas, así como las salidas de esta fase.

Tabla 7. Selección de los Factores de Posicionamiento

1. Selección de los Factores de Posicionamiento		
Entrada	**Herramientas y Técnicas**	**Salida**
- Artículos - Libros	- Revisión Literaria - Juicio de expertos	- Lista Factores

Fuente. *Autor*

Artículos

Recopilación de factores propuestos por artículos científicos, MOZ y Blogs especializados

Libros

Factores de posicionamiento recomendados por libros sobre SEO

Procedimiento:

Los factores internos más relevantes deben ser tomados mediante una exhaustiva revisión literaria de libros y artículos. También se pueden considerar aquellos recomendados por expertos y la propia experiencia. Es importante que el peso o valor de los factores pueda obtenerse mediante el rastreo de documentos HTML

Salida

La lista debe contener las siguientes cabeceras:

Tabla 8. Ejemplo de Lista de Factores

ID	Factor	Descripción	Tipo de Dato	Valores

4.2.2 Fase 2. Selección de palabras clave

Miles de millones de personas en todo el mundo realizan búsquedas todos los días mediante el envío de palabras clave. Una palabra clave es la

consulta textual enviada por los usuarios a los motores de búsqueda con la finalidad de obtener páginas web relevantes a su consulta (Aregay, 2014).

En esta fase se seleccionan las palabras clave más relevantes; para ello, se debe identificar la temática o nicho relacionado a la página web que se desea posicionar y elegir una palabra clave objetivo para ejecutarla en la herramienta de la fase 3. La tabla 9 muestra las entradas, herramientas y técnicas, así como las salidas de esta fase.

Tabla 9. Fase 2. Selección de las palabras clave

2. Selección de las palabras clave		
Entrada	**Herramientas y Técnicas**	**Salida**
- Palabra Clave Objetivo	- Keyword Planner de Adwords	- Lista de Palabras Clave

Fuente. *Autor*

Palabra Clave Objetivo

Es una palabra o frase principal el cual se desea posicionar en el motor de búsqueda, esta palabra puede ser genérica y orientada a un nicho. También cuenta con palabras clave relacionadas.

Keyword Planner de Adwords

Es la herramienta de investigación de palabras clave de Google, brinda estadísticas acerca de la frecuencia con que se buscan las palabras clave relacionadas a productos o servicios.

Procedimiento:

Se debe elegir una palabra clave objetivo en un único idioma y consultarla en la herramienta de palabras clave de Google (Keyword Planner de Adwords). Esta herramienta mostrará una lista de palabras clave relacionadas y ordenadas por volumen de búsquedas. Es recomendable que

la lista de palabras clave tenga la mayor cantidad posible de aquellas que se le relacionen

4.2.3 Fase 3. Rastreo de contenido

Los rastreadores web o las arañas se utilizan principalmente para recopilar diferentes tipos de información de las páginas web para su posterior procesamiento. También se pueden usar para automatizar las tareas de mantenimiento en una página web, como verificar enlaces, validar el código HTML o extraer textos específicos (Lavania, 2013).

En esta fase se deben extraer los valores de los factores de las páginas web. Para ello, se requiere de dos herramientas que serán ejecutadas en dos etapas: la primera se encarga de rastrear el contenido de resultados del motor de búsqueda por cada palabra clave, para después obtener las URL de las páginas indexadas; la segunda rastrea el contenido de las páginas web obtenidos en la primera etapa, esto con el fin de obtener los valores de cada factor interno. La tabla 10 muestra las entradas, herramientas y técnicas, así como las salidas de esta fase.

Tabla 10. Fase 3. Rastreo de Contenido

3. Rastreo de Contenido		
Entrada	**Herramientas y Técnicas**	**Salida**
- Lista de palabras clave - Lista de factores	- Rastrear el contenido de los resultados del motor de búsqueda - Rastrear el contenido de las páginas web indexadas	- Conjunto de datos de factores

Fuente. *Autor*

Lista de factores

Proviene de la fase 1

Lista de palabras clave

Proviene de la fase 2

Procedimiento:

En esta fase se debe ejecutar una herramienta para rastrear el contenido de los resultados del motor de búsqueda. Esta herramienta debe tomar cada palabra clave de la lista inicial, realizar la consulta en el motor de búsqueda, tomar las URL de las tres primeras posiciones de la primera página de resultados y tres de la quinta o mayor página de resultados. Para Ochoa (2012), los usuarios de Google solo visitan los primeros resultados antes de cambiar sus consultas, solo un 16 % de ellos pasa a la segunda página de resultados y menos del 1 % llegan hasta la cuarta página, por lo tanto, la justificación para tomar estas URL se debe a que las páginas web con mejor visibilidad y posicionamiento se encuentran en las 3 primeras posiciones; y las que se encuentran en la quinta página en adelante son las que prácticamente no tienen visibilidad y se puede considerar como páginas web no posicionadas. A mayor distancia entre las páginas posicionadas y no posicionadas, mayor será diferencia de los valores de los factores. Las URL a tomar únicamente deben ser documentos HTML. También debe considerar que para ejecutar el rastreador la sesión de usuario de la cuenta de Google debe estar cerrada, se debe ejecutar en un solo dominio de Google (p. ej., www.google.com.pe) y tomar únicamente las URL de los resultados orgánicos.

En la segunda etapa del rastreo del contenido de las páginas web indexadas, se debe tomar las URL obtenidas en la primera etapa y extraer los valores de cada factor de las páginas web, según la lista de factores obtenida en la fase 1. Las cabeceras del conjunto de datos de factores deben ser de la siguiente manera:

Tabla 11. Ejemplo del conjunto de datos de Factores

Posición	Factor1	Factor2	Factor3	…	FactorN
1	V	V	12		F
2	V	V	15		
3	F	F	16		V
51	V	V	12		F
52	V	V	15		V
53	V	V	14		V

Fuente. *Autor*

Se recomienda ejecutar esta fase durante un período superior a siete días con el fin de obtener una gran cantidad de datos.

4.2.4 Fase 4. Preparación de datos

En esa fase se prepara el conjunto de datos de los factores obtenidos en la fase 3, esto con el fin de tener datos completos y sin duplicidad. También se deben numerar normalizar o discretizar buscando adaptar los datos a las necesidades de los algoritmos. La tabla 12 muestra las entradas, herramientas y técnicas, así como las salidas de esta fase.

Tabla 12. Fase 4. Preparación datos

Preparación de Datos		
Entrada	**Herramientas y Técnicas**	**Salida**
Conjunto de datos de factores	Limpieza de datos Transformación de datos	Conjunto de datos de factores normalizados

Fuente. *Autor*

Conjunto de datos de factores

Proviene de la fase 3

Procedimiento:

- Eliminar registros duplicados e incompletos.
- Transformar todos los datos numéricos a escalas de ordinales como “Muy Alto, Alto, Normal, Bajo, Muy Bajo”.
- Transformar los valores booleanos a verdadero (V) y Falso (F).
- Transformar los valores del campo “Posición” en “Posicionado” para las 3 primeras posiciones y “NoPosicionado” para las siguientes 3 posiciones según el conjunto de datos de factores.

4.2.5 Fase 5. Aplicación de técnica de Machine Learning

Esta es la etapa en la que se genera la base de conocimiento aplicando una técnica de Machine Learning que genere reglas con base en el conjunto de datos de factores normalizados obtenidos en la fase 4, con el fin de extraer reglas de decisión que nos permitan proponer factores importantes para posicionar una página web. La tabla 13 muestra las entradas, herramientas y técnicas, así como las salidas de esta fase:

Tabla 13. Fase 5. Aplicación de técnica de Machine Learning

4. Aplicación de técnica Machine Learning		
Entrada	**Herramientas y Técnicas**	**Salida**
- Conjunto de datos de factores normalizados	- Herramienta para Machine Learning - Técnica de Machine Learning	- Conjunto de reglas

Fuente. *Autor*

Conjunto de datos de factores normalizados

Proviene de la fase 4

Herramienta para Machine Learning

Software libre o de pago que permita aplicar técnicas de machine Learning. Weka, Data Mining, Sciki-learn, etc

Técnica de Machine Learning

Técnica que genere reglas de decisión como árbol de decisiones, Random Forest, etc

Procedimiento:

- Seleccionar una herramienta para Machine Learning.
- Aplicar una técnica de Machine Learning que genere reglas de decisión.
- Las reglas generadas por la herramienta deben ser trasladadas a un conjunto de datos en forma de vector.
- Los factores de cada regla que no tengan valor se pueden reemplazar con una "X", este representa a cualquier valor, es decir, que su valor no influye en la decisión final, pero es recomendable aplicarlas.
- Considerar solo las reglas con la clasificación "Posicionado".

El conjunto de reglas generado sería el siguiente:

Tabla 14. Ejemplo del conjunto de Reglas

Posición	Factor1	Factor2	Factor3	…	Factor N
Posicionado	V	F	X	…	Alto
Posicionado	X	F	V	…	Bajo
NoPosicionado	V	V	F		X

Fuente. *Autor*

4.2.6 Fase 6. Recomendación de Factores de posicionamiento

El conjunto de reglas generado en la fase 6 contiene todas las reglas útiles para posicionar una página web. Según la página web a posicionar, se debe realizar un proceso de comparación hasta obtener la regla ideal y personalizada para la página web. La tabla 15 muestra las entradas, herramientas y técnicas, así como las salidas de esta fase:

Tabla 15. Fase 6. Recomendación de factores

6. Recomendación de factores		
Entrada	**Herramientas y Técnicas**	**Salida**
- Conjunto de Reglas - Página Web a posicionar	- Extraer valores de los factores de la página web a posicionar - Comparar las reglas con los valores de la página web	- Lista de factores recomendados

Fuente. *Autor*

Conjunto de Reglas

Proviene de la fase 5

Página Web a posicionar

Página web que se desea posicionar y extraerán las métricas de sus factores con el fin de compararlas con el conjunto de reglas.

Procedimiento:

- Mediante un rastreador de documentos, extraer los valores de los factores de la página web que se desea posicionar.
- Comparar cada registro del conjunto de reglas con los valores de los factores de la página web hasta encontrar el más similar, empezando con los factores de la página web que tengan el valor “Verdadero” y el menor número de registros filtrados, hasta llegar al mínimo de reglas.
- En caso de que hubiese más reglas, comparar con el tamaño del documento, número de enlaces externos y número de enlaces internos (el orden es con base en el menor número de registros filtrados).
- Seleccionar la regla que tenga menos factores a cambiar.

Finalmente, realizar los cambios respectivos, según los factores recomendados, en la página web

CAPÍTULO 5: EL SOFTWARE

El presente capitulo detalla la automatización del método propuesto mediante el desarrollo de un software, este tiene como entradas las palabras clave y como salida los factores de posicionamiento personalizados. También se describe la arquitectura y los módulos del software.

5.1 Propuesta de automatización

El método se divide en dos etapas, la primera consiste en extraer las métricas de los factores de posicionamiento de las páginas web indexadas al motor de búsqueda de Google de una temática determinada, pre-procesar y normalizar los datos, aplicar una técnica de Machine Learning y generar las reglas de decisión como base de conocimiento. La segunda etapa consiste en recomendar los factores de posicionamiento a un sitio web determinado mediante la comparación de sus métricas y las de las reglas. La siguiente figura muestra el procedimiento descrito.

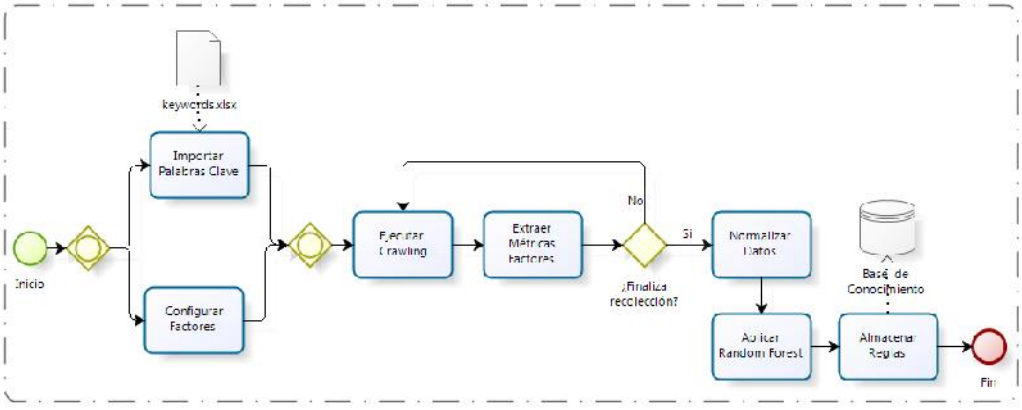


Figura 15. Proceso de generación de base de conocimiento

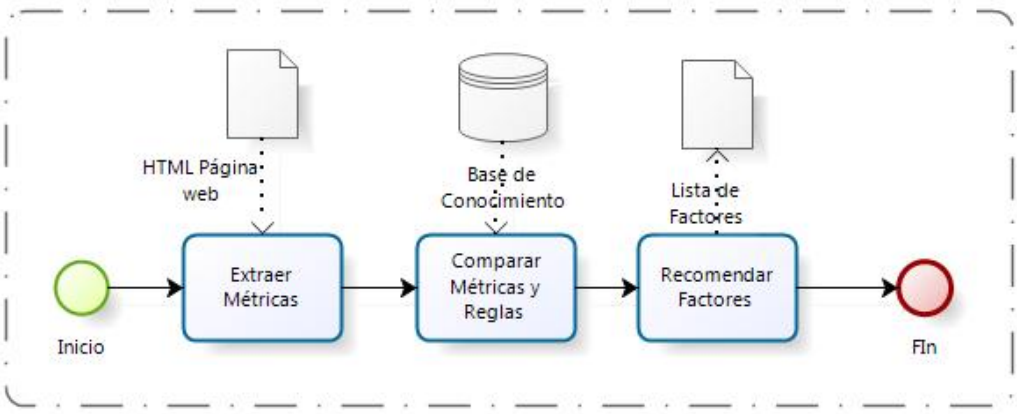


Figura 16. Proceso de recomendación de factores de posicionamiento.

Para automatización se considerarán los dos procesos, por lo cual se definió el siguiente requerimiento según la siguiente figura.

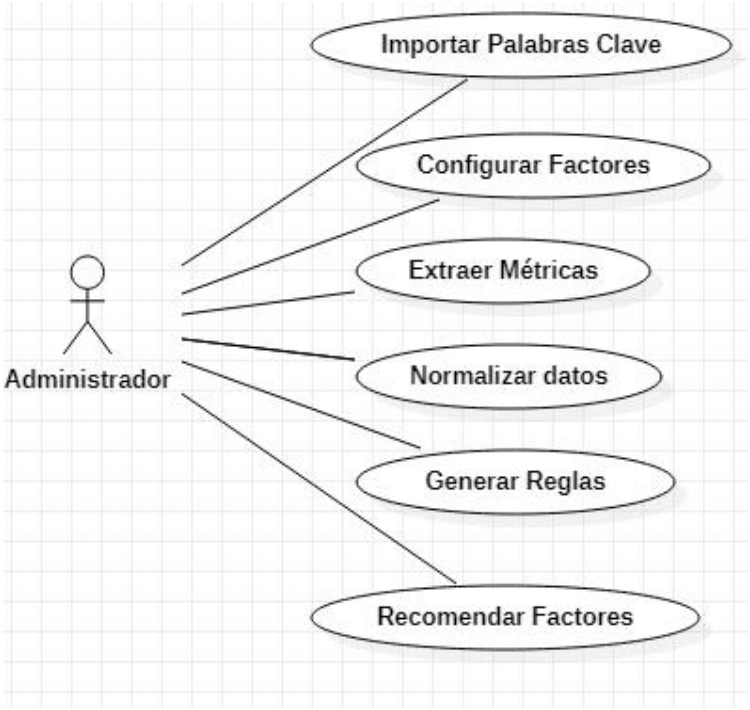


Figura 17. Diagrama de casos de uso

En consecuencia el flujo de la automatización del método es el siguiente:

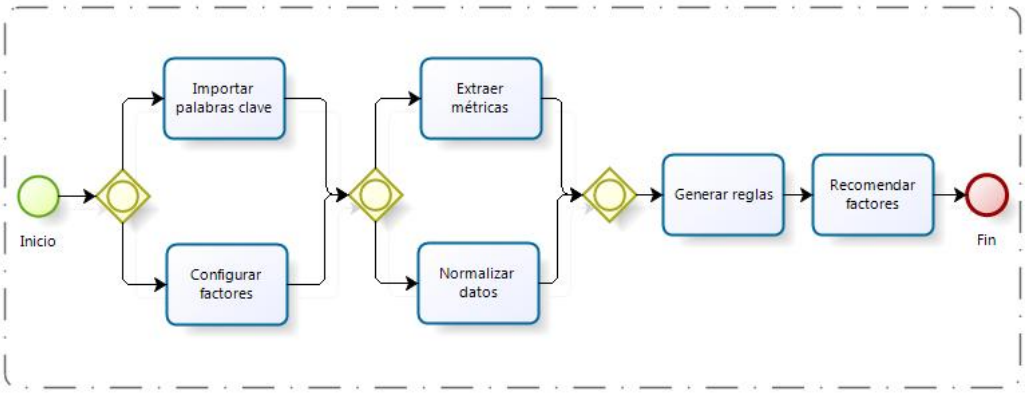


Figura 18. Propuesta de automatización

Según la figura 18 las funcionalidades del sistema son:

- **Importar palabras clave**. Se sube un archivo en excel con la lista de palabras claves a ejecutar en el motor de búsqueda de Google
- **Configurar factores**. Se configuran los factores que se desean rastrear, se crea la función con la lógica para extraer la métrica de cada factor
- **Extraer métricas**. Ejecuta un crawling encargado de extraer a los sitios web indexados en Google y extrae los valores o mpetricas de cada factor
- **Normalizar Datos.** Depura datos, limpia y transforma datos numéricos a ordinales.
- **Generar Reglas**. Ejecuta el algoritmo de random forest para generar los arboles y estos son pasados a vector para ser guardados en la base de datos

- **Recomendar factores**. Extrae las métricas de los factores de la página web a posicionar y las compara con las reglas de la base de conocimiento hasta obtener la regla que recomiende los factores personalizados para la página web

5.2 Descripción del Sistema

RANKING FACTOR es el nombre del software desarrollado para rastrear los resultados del motor de búsqueda de Google, los documentos de las páginas web, extrae las métricas de los factores de posicionamiento

internos, normaliza, aplica Random Forest y recomienda los factores necesarios para un determinado sitio web. La entrada de datos es el documento de Excel de palabras clave obtenidas y la salida un reporte de los factores personalizados para el sitio web. El esquema general del software se ilustra en la siguiente figura:

Figura 19.Esquema general del sistema

Fuente. Elaboración Propia

Requerimientos funcionales

Tabla 16.Lista de requerimientos funcionales del sistema

Id	Nombre	Descripción
RF001	IMPORTAR PALABRAS CLAVE	El sistema debe permitir la importación de la lista de palabras clave en .xlsx
RF002	EXTRAER URL	El software deberá extraer las URL de los resultados del buscador de Google por cada palabra clave
RF003	GUARDAR DOCUMENTOS	El software descarga los todos documentos HTML de cada URL y los debe almacenar en carpetas nombradas por fecha
RF004	INTEGRAR	El software debe permitir integrar ficheros .php

	ARCHIVOS	y la lista de factores conteniendo el procedimiento extraer la métrica de cada factor de posicionamiento de los documentos
RD005	NORMALIZAR DATOS	El software normaliza los datos permitiendo al usuario seleccionar los valores que deben ser convertidos a valores ordinales
RF006	EJECUTAR RAMDOM FOREST	El software debe ejecutar el algoritmo Random Forest y generar los arboles de decisión
RF007	CONVERTIR ARBOLES A VECTOR	El software transforma la estructura de árbol en vector y los almacena en la base de datos
RD008	REPORTE DE FACTORES	El software extrae las métricas de un sitio web, compara con las reglas y emite un reporte de los factores de posicionamiento por cada documento y periodo

Fuente. Elaboración Propia

Requerimientos no funcionales

Tabla 17. Lista de los requerimientos no funcionales del software

Id	**Descripción**
RNF001	Disponibilidad 24/7
RNF002	Los datos deberán almacenarse en la base de datos y los documentos en una estructura de directorios
RNF003	Integridad de información
RNF004	El software deberá tener licencia libre
FNF005	Portabilidad
RNF006	Lenguaje de programación PHP 7.2
RNF007	Base de datos Mysql 6
RNF008	Apache 2.0

Fuente. Elaboración Propia

5.3 Arquitectura del Software

Diagrama de Despliegue

A continuación se muestra la la disposición fisica de los artefactos del software

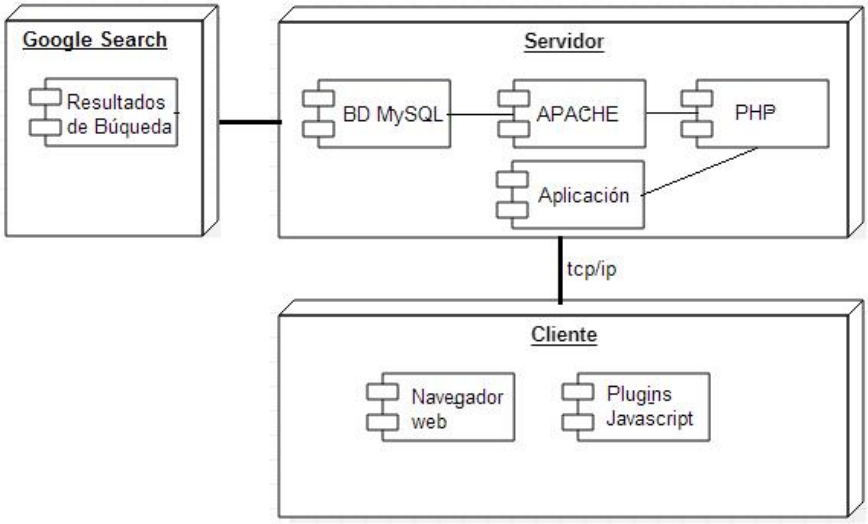


Figura 20. Modelo de diagrama de despliegue

Fuente. Elaboración Propia

Modelo de Componentes

La división en componentes y las dependencias entre estos de software de muestra en la siguiente figura

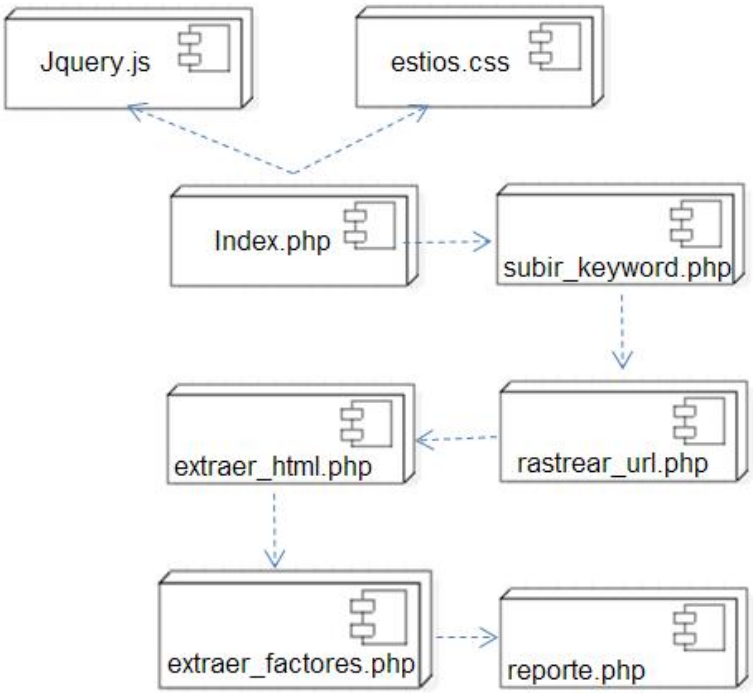


Figura 21. Modelo de diagrama de componentes

Fuente. Elaboración Propia

Requerimientos de Hardware

Los requerimientos mínimos de hardware que se requieren para el funcionamiento dela aplicación es la siguiente:

PC cliente

Procesador Intel core 2 duo

Velocidad: 1.06Ghz

RAM: 4Gb

Servidor

Intel core i7

Velocidad 3.4Ghz

RAM: 32gb

Disco 1Tb

Requerimientos de Software

Los requerimientos mínimos de software son los siguientes:

Para el desarrollo del software

- SublimeText 3.0
- XAMMP
- Jquery

Para el uso del software

- Software Operativo windows 7
- Navegador Google Chrome 54.x

5.4 Módulos del software

Módulo Importar Palabras Clave

Este módulo permite cargar el archivo Excel que contiene una lista de palabras clave tomadas de la herramienta de palabras clave de Google.

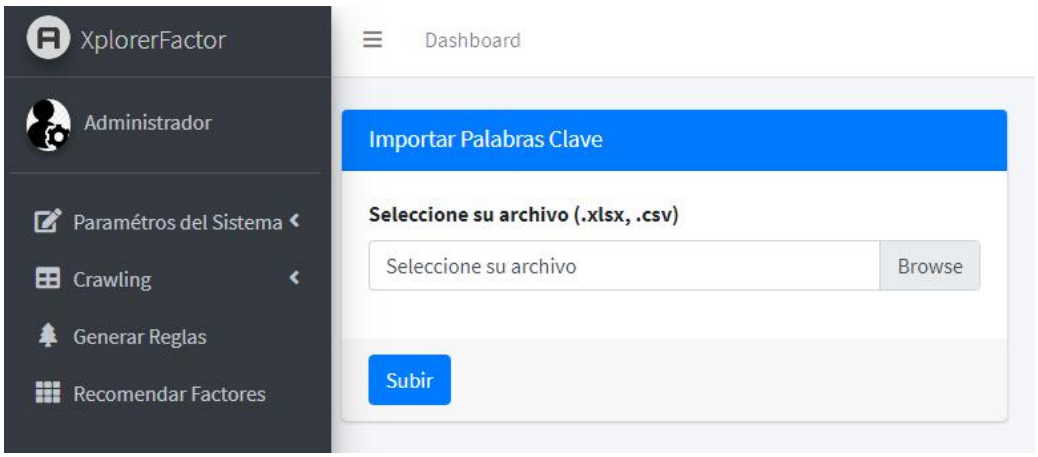


Figura 22. Módulo importar palabras clave

Fuente. Elaboración Propia

Tabla 18. Ejemplo de estructura del archivo Excel

ID	Palabra Clave
1	Turismo Perú
2	Paquetes turísticos
3	Tours Perú
4	Viajes

Fuente. Elaboración Propia

Módulo Configurar Factores

En este módulo se agregan los factores internos que se desean analizar y extraer, para ello el usuario debe asignarle un nombre, descripción y un algoritmo que se encargue de extraer la métrica. La entrada que debe usar es $HTML->find() y la salida $VALOR

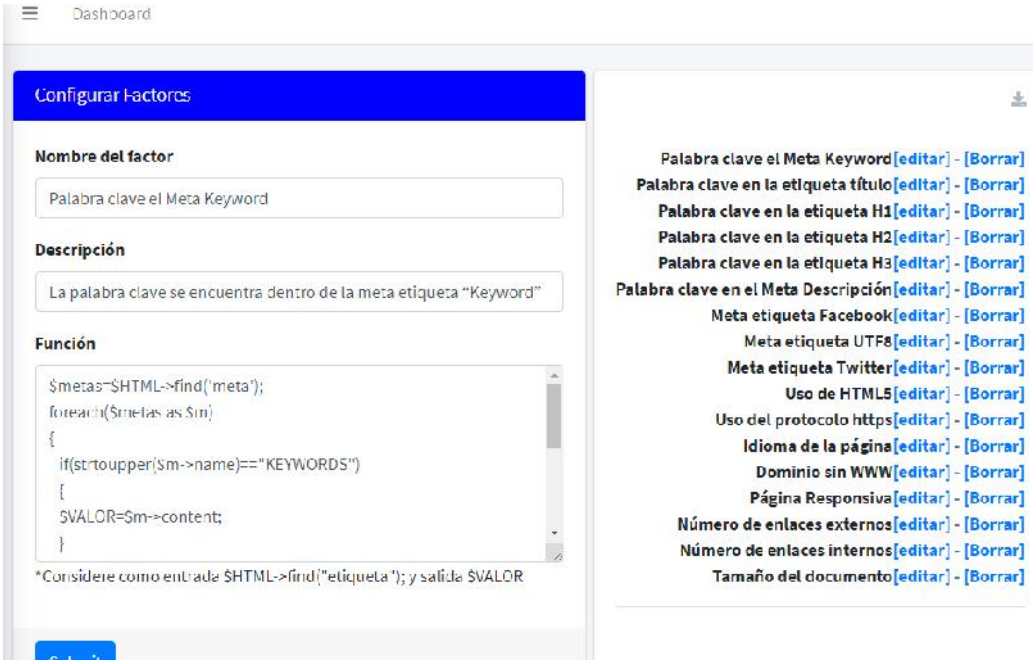


Figura 23. Módulo Configurar Factores

Módulo Extraer Métricas

Este módulo ejecuta consultas al motor de búsqueda de Google por cada palabra clave del Excel cargado anteriormente y por cada consulta ejecutada se extraen las URL de las primeras posiciones y las que se encuentran en la página 5

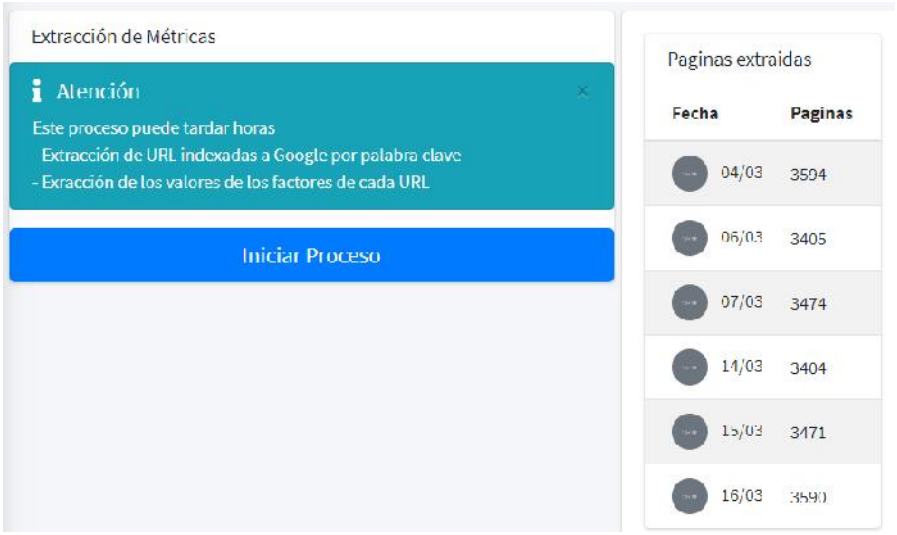


Figura 24. Módulo Extraer Métricas

Fuente. Elaboración Propia

Módulo Normalizar Datos

En este módulo se realiza la limpieza de datos incompletos, duplicados y se transformar las unidades de medida según los parámetros del usuario

Figura 25.Módulo Normalizar Datos

Tabla 19.Ejemplo de los datos normalizados por el software

Posición	k_title	k_url	k_meta_k	TamañoWeb	facebook	k_h1
Posicionado	F	F	F	Alto	F	F
NoPosicionado	F	F	F	Bajo	F	F
NoPosicionado	F	F	F	Alto	V	F
Posicionado	F	F	F	MuyAlto	V	F

Posicionado	F	V	F	Alto	V	F
Posicionado	F	F	F	Alto	V	F
NoPosicionado	F	F	F	Bajo	V	F
NoPosicionado	V	F	V	Alto	V	V
NoPosicionado	F	F	F	MuyBajo	V	F
Posicionado	F	F	F	Bajo	V	F
Posicionado	F	F	F	MuyAlto	F	F
Posicionado	F	F	F	Bajo	V	F
NoPosicionado	F	F	F	MuyBajo	V	F
Posicionado	F	F	V	Bajo	V	F
Posicionado	F	F	F	Alto	F	F
Posicionado	F	F	F	Alto	F	F

Fuente. Elaboración Propia

Módulo Generar Reglas

En este módulo se ejecuta el algoritmo de Random Forest, previamente se revisaron diferentes algoritmos obteniendo en este un mayor porcentaje de precisión. También se transforman los arboles generados a vector para poder ser almacenados en la base de datos

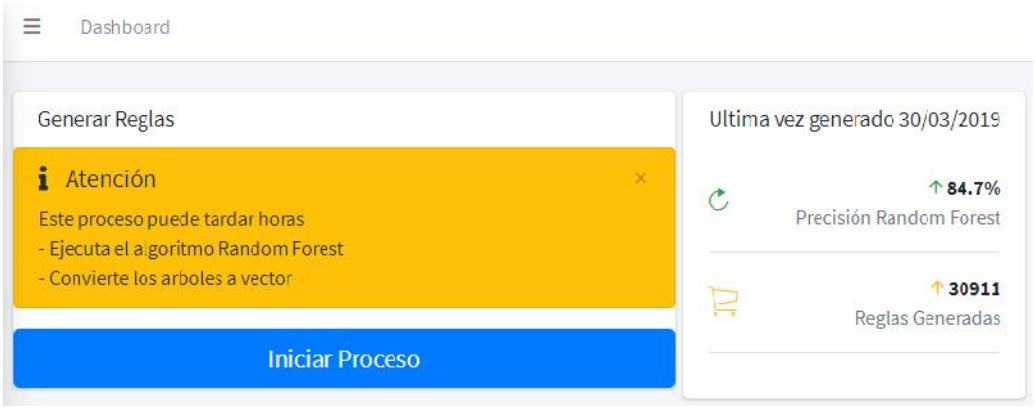


Figura 26. Módulo Generar Reglas

Módulo Recomendar Factores

En este módulo se ingresa la URL de la página web a posicionar y la palabra clave objetivo, el software extrae las métricas del la pagina web y las

compara con las reglas almacenadas en la base de datos. Según los diferentes filtros durante la comparación finalmente el software recomienda los factores necesarios para que la página web pueda mejorar su posicionamiento.

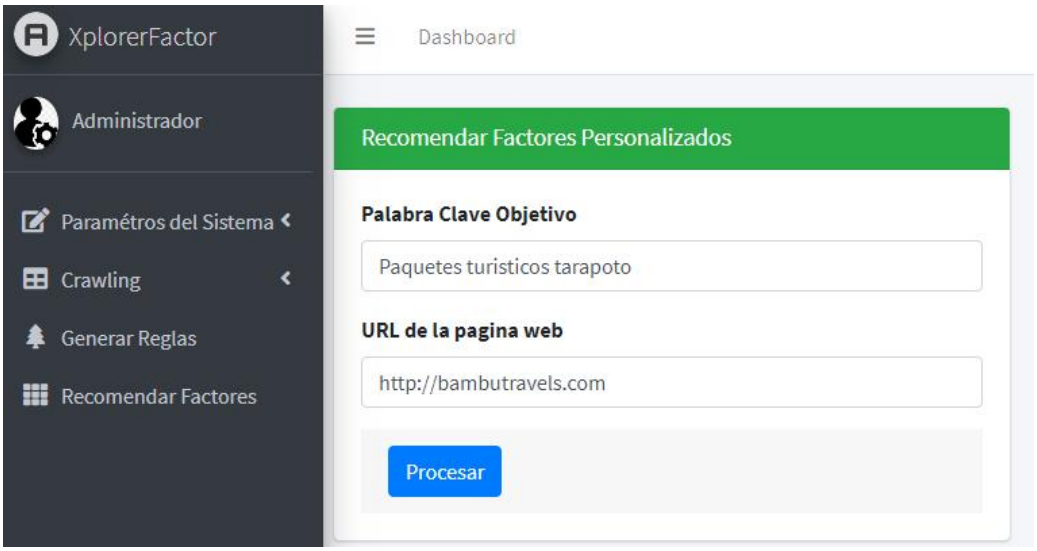


Figura 27. Módulo Recomendar Factores

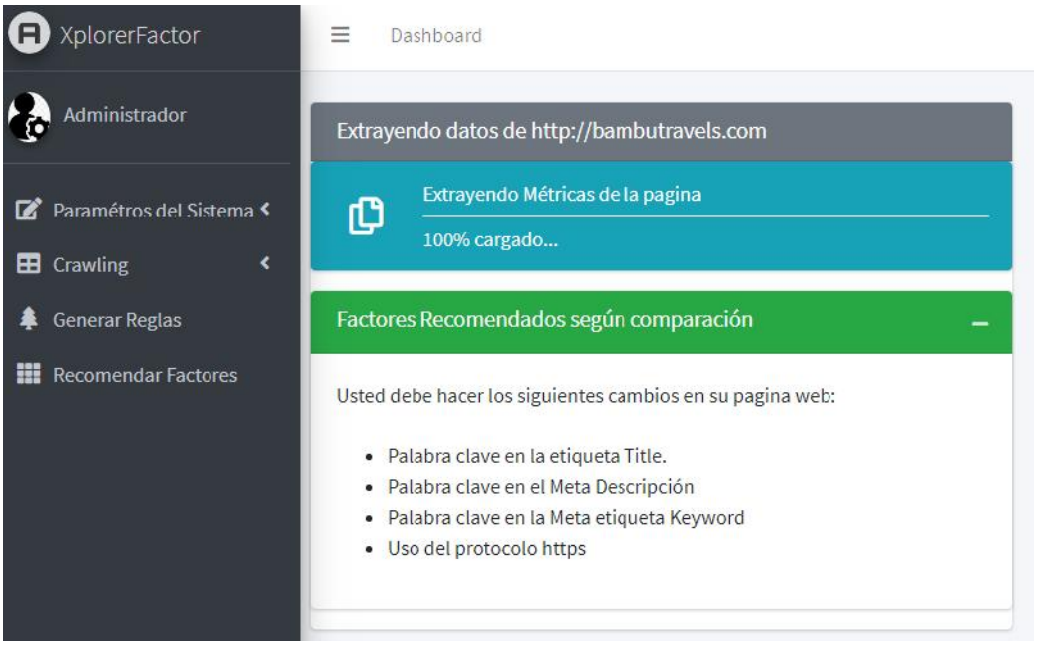


Figura 28.Módulo con los factores recomendados

5.5 Pruebas de Software

Se ejecutaron una serie de pruebas básicas para validar el funcionamiento del software

Tabla 20. Pruebas del software ejecutadas

Id	Prueba	Estado
1	Verificar que se cargan todas las palabras clave	Correcto
2	Verificar que se extraen las URL por cada consulta	Correcto
3	Verificar que solo se recuperan documentos HTML por cada URL	Correcto
4	Verificar que las funciones extraen los factores correctos por documento	Correcto
5	Verificar que la normalización de datos	Correcto

Fuente. Elaboración Propia

CAPÍTULO 6: VALIDACION

El presente capitulo consiste en la validación del método propuesto, para ello se seleccionará una página web sin visibilidad en una determinada palabra clave y se aplicará el método propuesto (con apoyo del software desarrollado) con el fin de posicionarlo y aumentar la visibilidad en el motor de búsqueda de Google.

6.1 Caso de estudio

Para validar se seleccionó bambutravels.com, página web que pertenece a una agencia de viajes peruana dedicada a la venta de vuelos y paquetes turísticos en la ciudad de Tarapoto, Perú. Sus principales clientes son turistas nacionales.

La palabra clave objetivo seleccionada, con la que se busca posicionar la página web, es "paquetes turisticos a Tarapoto". Esta palabra clave cuenta con 112,000 resultados en google.com.pe y, según la herramienta de Keyword Planner de Adwords, mantiene un promedio de 120 búsquedas mensuales con un nivel de competencia alto. No se considera el uso de la tilde puesto que los usuarios no suelen utilizar los acentos en sus consultas.

Antes de aplicar el método, la página web se encontraba en la posición 46 en los resultados de búsqueda en Google.com.pe y según Ochoa (2012), los usuarios no llegan hasta la cuarta página de los resultados de búsqueda, lo cual implica que la página web no tiene visibilidad en el motor de búsqueda de Google.

6.2 Aplicación del método de posicionamiento

A continuación se detalla la aplicación del método propuesto por fases y los resultados obtenidos.

Fase 1. Selección de los Factores de Posicionamiento

Para este caso de estudio se han considerado 18 factores de posicionamiento internos que fueron tomados por medio de una revisión

literaria y de la experiencia de los autores. Los factores seleccionados fueron los siguientes:

Tabla 21. Lista de Factores Seleccionados

ID	Factor	Descripción	Tipo de Dato	Valores
1	Palabra clave en la URL	La palabra clave se encuentra en la URL de la página	Booleano	V o F
2	Palabra clave el Meta Keyword	La palabra clave se encuentra dentro de la meta etiqueta “Keyword”	Booleano	V o F
3	Palabra clave en la etiqueta título	La palabra clave se encuentra dentro de la meta etiqueta “title”	Booleano	V o F
4	Palabra clave en la etiqueta H1	La palabra clave se encuentra dentro de la meta etiqueta “h1”	Booleano	V o F
5	Palabra clave en la etiqueta H2	La palabra clave se encuentra dentro de la meta etiqueta “h2”	Booleano	V o F
6	Palabra clave en la etiqueta H3	La palabra clave se encuentra dentro de la meta etiqueta “h3”	Booleano	V o F
7	Palabra clave en el Meta Descripción	La palabra clave se encuentra dentro de la meta etiqueta “description”	Booleano	V o F
8	Meta etiqueta Facebook	Hace uso de las meta etiquetas para Facebook	Booleano	V o F
9	Meta etiqueta UTF8	Hace uso de la meta etiqueta utf8 para la codificación de la página	Booleano	V o F
10	Meta etiqueta	Hace uso de las metas	Booleano	V o F

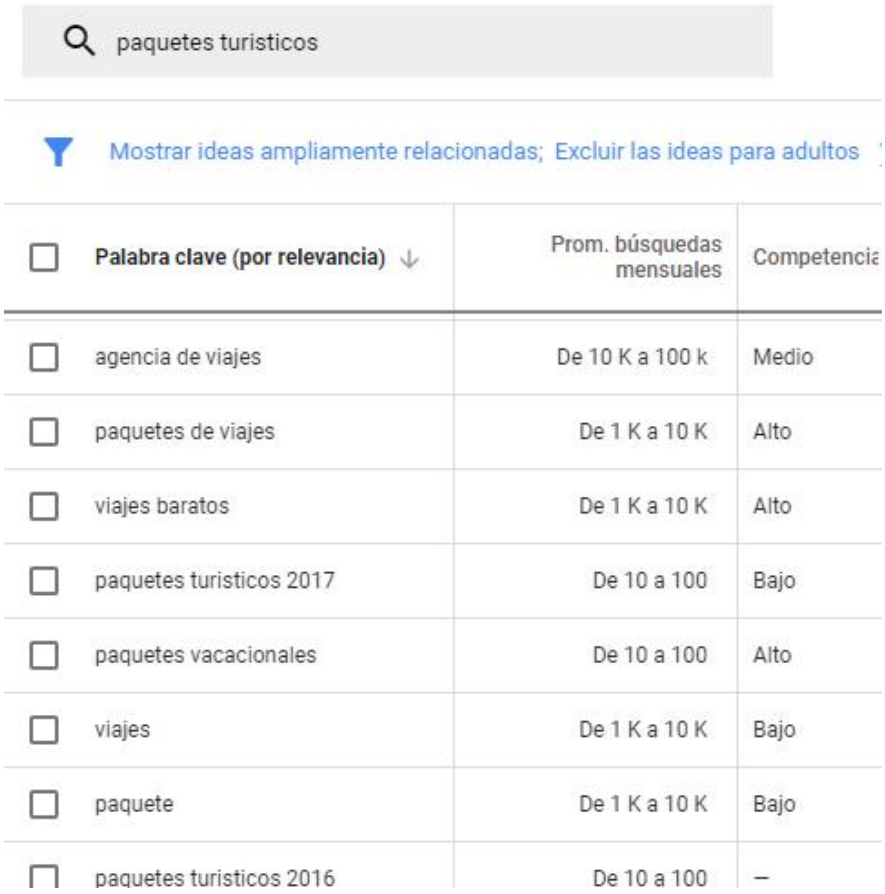
	Twitter	etiquetas para Twitter		
11	Uso de HTML5	La página web esta en HTML	Booleano	V o F
12	Uso del protocolo https	La página web usa un certificado de seguridad SSL	Booleano	V o F
13	Idioma de la página	Se especifica el idioma de la página	Booleano	V o F
14	Dominio sin WWW	Si el dominio esta indexado sin “www”	Booleano	V o F
16	Página Responsiva	La página web es responsivo	Booleano	V o F
16	Número de enlaces externos	Número total de enlaces externos	Numérico	>=0
17	Número de enlaces internos	Número total de enlaces internos	Numérico	>=0
18	Tamaño del documento	Tamaño del documento HTML de la página web	Numérico	>0

Fuente. *Autor*

Estos factores fueron agregados al software con su respectiva función

Fase 2. Selección de las Palabras Clave

Debido a que la página web a posicionar está dentro del nicho turismo, la palabra clave objetivo elegida fue “paquetes turísticos”. A partir de esto y con el uso de la herramienta Keyword Planner de Adwords se tomaron 599 palabras clave relacionadas a la palabra clave objetivo. Con ello se generó la lista con 600 palabras clave.



Palabra clave (por relevancia)	Prom. búsquedas mensuales	Competencia
agencia de viajes	De 10 K a 100 k	Medio
paquetes de viajes	De 1 K a 10 K	Alto
viajes baratos	De 1 K a 10 K	Alto
paquetes turisticos 2017	De 10 a 100	Bajo
paquetes vacacionales	De 10 a 100	Alto
viajes	De 1 K a 10 K	Bajo
paquete	De 1 K a 10 K	Bajo
paquetes turisticos 2016	De 10 a 100	–

Figura 29. Ideas de palabras clave

Fuente. Keyword Planner de Adwords

Las 600 palabras clave se pasaron a un archivo de Excel para luego subirlas al software

Fase 3. Rastreo de Contenidos

Se desarrolló un software para rastrear los resultados de búsqueda del motor de búsqueda de Google por cada palabra clave consultada (lista de palabras clave) y, posteriormente, se rastrearon los documentos HTML, de cada página web, con el fin de extraer los valores de cada factor interno, según la lista de factores definida en la primera fase. El módulo del software que se encargan de este procedimiento es “Extraer Métricas”

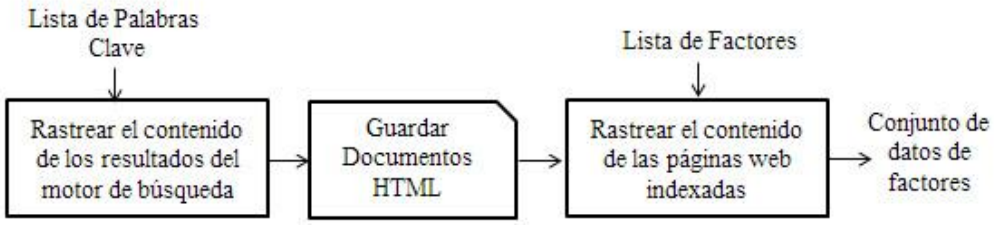


Figura 30. Proceso de rastreo

Fuente. *Elaboración Propia*

Durante un determinado período, mediante el software, se procedió a extraer las URL de los resultados del motor de búsqueda de Google por palabra clave enviadas y sus métricas, para ello, se han considerado los siguientes filtros:

- Solo búsquedas en idioma español
- Búsquedas sin la sesión activa de la Cuenta de Google
- Búsquedas al dominio Google.com.pe
- Solo se consideran los resultados orgánicos
- Se tomaron las URL de las posiciones 1,2,3 y 51,52,53

La solicitud ejecutada por la herramienta fue a la siguiente URL: https://www.google.com.pe/search?q=**{Palabra Clave}&**lr=lang_es

La herramienta ejecutó la lista de palabras clave en el motor de búsqueda de Google durante 7 días. En total se extrajeron 3594 URL por día, lo cual hizo un total de 25,158 URL almacenados en la base de datos. Posteriormente, se almacenó en un directorio todos los documentos por cada URL válida (no se incluyó documentos diferentes a HTML), por lo que se obtuvo un Dataset de factores con 19,385 registros.

Fase 4. Preparación de datos

En esa fase se normalizaron los valores de los factores “Número de enlaces externos”, “Número enlaces internos” y “Tamaño del documento” a escala ordinal basados en la distribución de Gauss. En el modulo de “Normalizar

Datos" del software se parametrizaron estos factores para que sean convertidos a ordinal.

links_externos

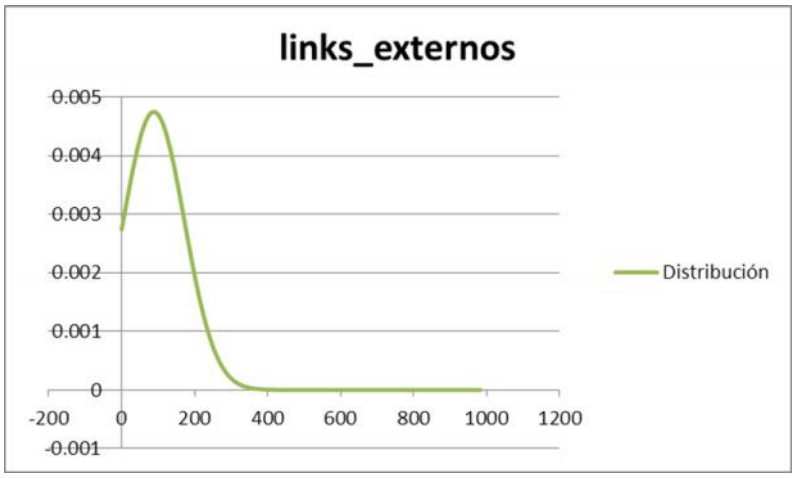


Figura 31. Campana de Gauss de enlaces externos
Fuente. Elaboración Propia

Según la campana se definió la siguiente distribución para los enlaces externos

Tabla 22. Tabla de distribución Enlaces Externos

Nro de enlaces	**Escala**
[>125]	MuyAlto
[75 - 125]	Alto
[50 - 75]	Normal
[25 - 75]	Bajo
[0 - 25]	Muy Bajo

Fuente. *Elaboración Propia*

Links_Internos

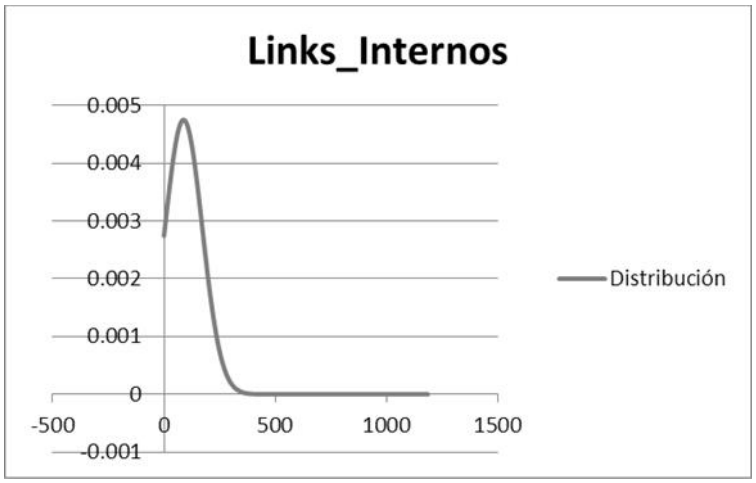


Figura 32. Campana de Gauss enlaces internos

Fuente. Elaboración Propia

Tabla 23. Tabla de distribución Enlaces Internos

Nro de enlaces	**Escala**
[>200]	MuyAlto
[100 - 200]	Alto
[50 - 100]	Normal
[50 - 25]	Bajo
[0 - 25]	Muy Bajo

Fuente. Autor

Tamaño

El tamaño del documento esta expresado en Kilobytes

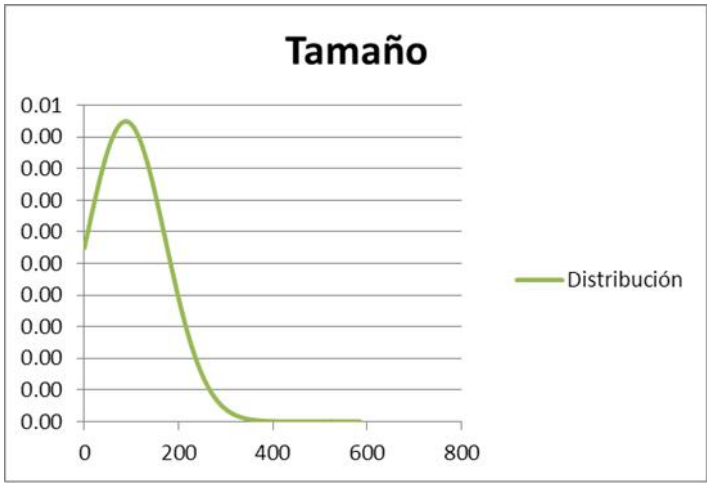


Figura 33. Campana de Gauss para Tamaño del documento

Fuente. Elaboración Propia

Según la campana se definió la siguiente distribución para los enlaces externos

Tabla 24. Tabla de distribución tamaño del documento

Tamaño	**Escala**
[>200]	MuyAlto
[150 - 200]	Alto
[100 - 150]	Normal
[50 - 100]	Bajo
[0 - 50]	Muy Bajo

Fuente. *Autor*

También se reemplazaron los valores “1, 2 ,3” del campo posición por “Posicionado” y los valores “51, 52 ,53” por “NoPosicionado”. Al hacer este cambio se generaron registros duplicados, los cuales fueron eliminados haciendo un conjunto de datos de 1576 registros.

Posicion	k_title	k_url	k_meta_k	links_externos	links_internos	googleplus	facebook	k_h1
Posicionado	F	F	F	Bajo	Muy Alto	F	F	F
NoPosicionado	F	F	F	Muy Bajo	Normal	F	F	F
NoPosicionado	F	F	F	Bajo	Muy Bajo	F	V	F
Posicionado	F	F	F	Muy Alto	Alto	V	V	F
Posicionado	F	V	F	Muy Alto	Muy Alto	F	V	F
Posicionado	F	F	F	Bajo	Muy Bajo	V	V	F
NoPosicionado	F	F	F	Alto	Normal	V	V	F
NoPosicionado	V	F	V	Muy Bajo	Alto	F	V	V
NoPosicionado	F	F	F	Alto	Muy Bajo	F	V	F
Posicionado	F	F	F	Alto	Muy Bajo	F	V	F
Posicionado	F	F	F	Bajo	Alto	F	F	F
Posicionado	F	F	F	Normal	Muy Bajo	V	V	F
NoPosicionado	F	F	F	Alto	Muy Bajo	V	V	F
Posicionado	F	F	V	Alto	Muy Alto	V	V	F
Posicionado	F	F	F	Bajo	Alto	F	F	F
Posicionado	F	F	F	Muy Bajo	Muy Bajo	F	F	F

Figura 34. Fragmento del DataSet de factores de posicionamiento final

Fuente. Elaboración Propia

Fase5. Aplicación de técnica de Machine Learning

Esta es la etapa se genera el modelo de conocimiento aplicando la técnica de Machine Learning con el software.

Para este caso se eligió el algoritmo de aprendizaje supervisado Random Forest con el fin de extraer reglas de decisión que nos permitan proponer factores importantes para posicionar una página web.

Para la validación, se utilizó la herramienta WEKA 3.8.2 con las siguientes configuraciones:

Número de árboles: Durante el entrenamiento con todos los registros se modificaron el número de árboles obteniendo una variación en la precisión como se muestra en la siguiente figura:

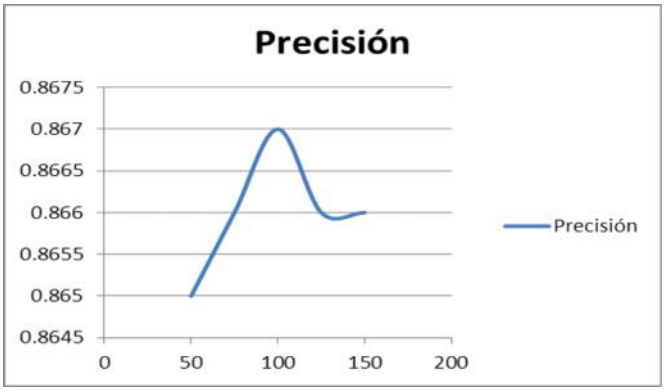


Figura 35. Precisión por número de árboles
Fuente. *Elaboración Propia*

Según los testeos realizados el número de árboles adecuado fue de 100 por tener una mayor precisión

Dimensión: 18 Atributos
Conjunto de entrenamiento: 1572 instancias
Número de árboles: 100
Regla de clasificación: Posicionado, NoPosicionado

Los resultados de WEKA mostraron una precisión promedio de 84.8 %. La tabla 23 muestra el resumen de los resultados:

Tabla 25. Resumen de resultados con WEKA

	Posicionado	**NoPosicionado**
Precisión	82.6 %	85.5 %
Reglas	30911	33358

Fuente. *Elaboración Propia*

Instancias correctamente clasificadas: 1337
Instancias incorrectamente clasificadas: 235

Media de precisión: 84.8 %
Número de reglas generadas: 64269
Matriz Confusión

```
  a    b   <-- classified as
219  189 |   a = Posicionado
 46 1118 |   b = NoPosicionado
```

Los arboles generados por Random Forest mediante WEKA se transformaron a vector.

Finalmente, se eliminaron las reglas duplicadas y se generó una lista de reglas únicamente con la regla de clasificación “Posicionado”, lo cual hizo un total de 30911 reglas.

Posicion	k_title	k_url	k_meta_k	links_eternos	links_internos	k_h1	k_h2	k_h3	k_meta_desc
posicionado				muybajo	bajo			V	
posicionado			V	muybajo	bajo			V	V
posicionado				muyalto	muyalto		F	V	
posicionado	F		V		muyalto	F			V
posicionado	F		V	normal	muybajo			v	
posicionado	V		V	normal	alto				V
posicionado			V	normal	bajo		F	V	
posicionado				bajo			F		V
posicionado		F		muyalto			F	V	V
posicionado							v	V	
posicionado	F	F			F		v		
posicionado				F	V		v		
posicionado							v		

Fase 6. Recomendación de factores

Como se mencionó anteriormente, la palabra clave con la que desea posicionar la página web bambutravels.com es “paquetes turisticos a Tarapoto”, cabe mencionar que los resultados del motor de búsqueda de Google varían el uso de la tilde, pero en este estudio no se ha considerado el uso de la tilde en la palabra clave debido a que los usuarios no suelen utilizarlas con frecuencia en sus consultas.

Mediante el software se rastreó la página bambutravels.com para obtener los valores de sus factores internos. Posteriormente, el software internamente

comparó los valores de la página web con la regla más similar que nos dio el método de la siguiente manera:

- El primer filtro: “Página Responsiva” con 7609 reglas
- El segundo filtro: “Uso de HTML5” con 2313 reglas
- El tercer filtro: “Tamaño del documento” con 464 reglas
- El cuarto filtro: “Número de enlaces externos” con 59 reglas
- El quinto filtro: “Número de enlaces internos” con 18 reglas

Debido a que el quinto filtro generó 18 reglas, se seleccionó la regla que menos cambios recomienda, así, quedó la siguiente tabla:

Tabla 26. Comparación de los factores de la página web y regla válida

ID	Factor	Valor	Regla	¿Requiere cambio?
1	Palabra clave en la URL	F	X	Opcional
2	Palabra clave en la Meta etiqueta Keyword	F	X	Opcional
3	Palabra clave en la etiqueta Title	F	X	Opcional
4	Palabra clave en la etiqueta H1	F	F	No
5	Palabra clave en la etiqueta H2	F	X	Opcional
6	Palabra clave en la etiqueta H3	F	X	Opcional
7	Palabra clave en la Meta etiqueta Description	F	X	Opcional
8	Meta etiqueta Facebook	F	X	Opcional
9	Meta etiqueta UTF8	F	X	Opcional
10	Meta etiqueta Twitter	F	X	Opcional
11	Uso de HTML5	V	V	No
12	Uso del protocolo https	F	V	Si
13	Idioma de la página	F	X	Opcional
14	Dominio sin WWW	F	X	Opcional
15	Página Responsiva	V	V	No
16	Número de enlaces externos	Muy	Bajo	No

17	Número de enlaces internos	Bajo(3) Muy Bajo (21)	Muy Bajo	Opcional
18	Tamaño del documento	Normal (58.7kb)	Normal	No

Fuente. Autor

Según la tabla 26, el método, mediante el software, reveló obligatoriamente el “Uso del protocolo https” y no usar “Palabra clave en la etiqueta H1”, además, no cambiar los factores “Uso de HTML5”, “Página Responsiva”, “Número de enlaces externos”, “Número de enlaces internos”, “Tamaño del documento”. La regla quedó de la siguiente manera:

Página Responsiva = V

| Uso de HTML5 = V

| | Tamaño del documento = Normal

| | | Número de enlaces externos = Normal

| | | | Número de enlaces internos = Muy Bajo

| | | | | Uso del protocolo https = V

| | | | | | Palabra clave en la etiqueta H1= F: Posicionado

Algunos factores con valor X fueron cambiados de manera opcional, no influyen en la regla, pero se recomienda utilizarlas.

Tabla 27. Cambios efectuados en la página web

Factor	**Cambios realizados**
Palabra clave en la etiqueta Title. (sin tilde)	<title>Paquetes Turisticos a Tarapoto 2019 - Agencia de Viajes en Tarapoto</title>
Palabra clave en el Meta	<meta name=’description’

Descripción (sin tilde)	content=’Ofrecemos Paquetes Turisticos a Tarapoto, tenemos paquetes todo incluido, tours, hoteles, vuelos. Disfruta semana Santa, fiestas patrias, semana larga’ >”
Palabra clave en la Meta etiqueta Keyword (sin tilde)	<meta name=’keywords’ content=’paquetes turisticos a tarapoto, paquetes turisticos, paquetes turisticos tarapoto, paquetes turisticos baratos, tours en tarapoto, agencia de viajes, operador turistico, viajes Tarapoto, paquete a Tarapoto’ >
Uso del protocolo https	Instalación de un certificado digital https://bambutravels.com

Fuente. Autor

Posteriormente, después de los cambios realizados, mediante la herramienta Google Search Console, se solicitó la re-indexación de la página web.

6.3 Discusión de Resultados

Para el seguimiento del posicionamiento de la página web https://bambutravels.com/, se utilizó la herramienta de rendimiento de Google Search Console. Esta herramienta provee a los propietarios de sitios web la oportunidad de comprobar el estado de indexación en el buscador y optimizar su visibilidad.

Después de los cambios realizados con base en las recomendaciones del método, la figura 36 muestra la evolución del posicionamiento con la palabra clave “paquetes turisticos a tarapoto” durante 3 meses, de este modo la

página web logró alcanzar la posición media máxima de 3.5 en los resultados del motor de búsqueda Google.com.pe cuando su posición inicial antes de aplicar el método era 46 lo cual significaba que no tenía visibilidad.

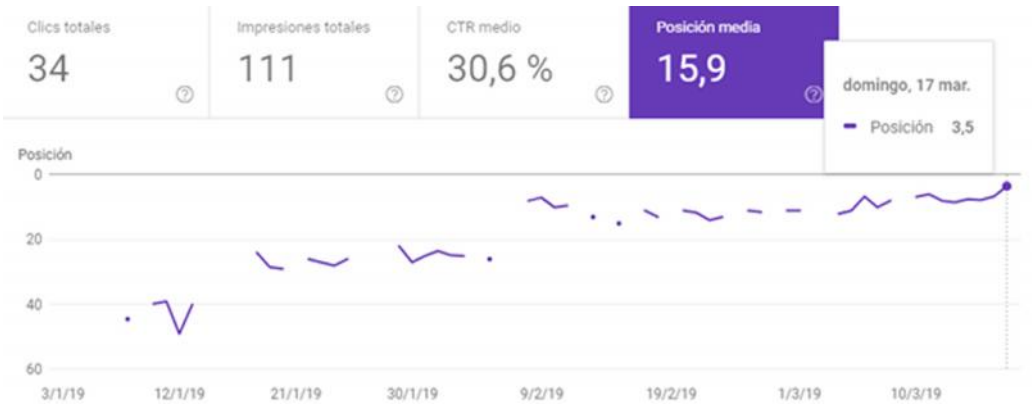


Figura 36. Evolución del posicionamiento con la palabra clave “paquetes turísticos a Tarapoto” en Google.com.pe después de aplicar las recomendaciones del método

Fuente. Google Search console

En general, sin considerar palabras clave específicas y zona geográfica, la figura 37 muestra que la posición media de la página web https://bambutravels.com/ fue de 16.8, siendo 98.4 la posición mínima(sin aplicar el método) y 5.6 la posición máxima (después de aplicar el método) alcanzada, lo cual indica que también existieron mejoras en el posicionamiento con otras palabras clave, las palabras clave más comunes fueron “tours tarapoto 2019“,“agencia de viajes bambu“,“bambu tours“,“paquetes turisticos tarapoto 2019“,“tarapoto paquetes turísticos“(con tilde) y “tours en tarapoto precios“

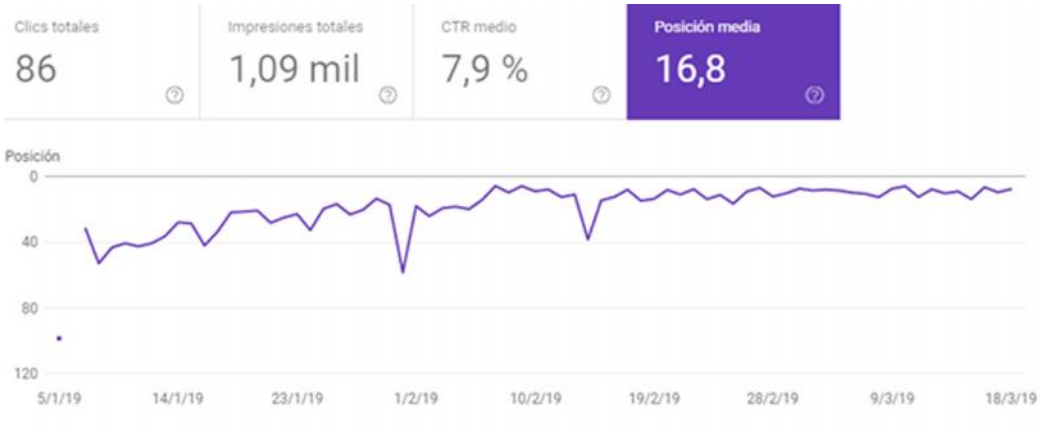


Figura 37. Evolución general del posicionamiento

Fuente. Google Search console

CAPÍTULO 7: CONCLUSIONES

7.1 Conclusiones

- Se elaboró un método para recomendar los factores de posicionamiento más relevantes en el posicionamiento web orgánico de forma personalizada, esto con la finalidad que una página web pueda alcanzar mejores posiciones en los resultados de búsqueda de Google
- Basados en la revisión sistemática de la literatura se identificaron los factores de posicionamiento internos que utilizaría el motor de búsqueda de Google para posicionar las páginas web
- Se aplicó la técnica de Random Forest para obtener las reglas que posicionen a una página web con un porcentaje de precisión de 84.8%.
- Se desarrolló un software para automatizar el método propuesto
- Se posicionó una página web, esta inicialmente no se tenía visibilidad en los resultados de búsqueda, pero, luego de los cambios recomendados por el método, esta logró mejoras significativas en su posicionamiento, así, alcanzó la primera página de los resultados del motor de búsqueda de Google con una posición media máxima de 3.5. Si bien es cierto la página web no alcanzó las tres primeras posiciones, los resultados conseguidos son optimistas, ya que se incrementó la visibilidad en el motor de búsqueda. Además de las palabra clave ya expuestas, también se mejoró su posicionamiento con otras palabras clave.
- El constante cambio del algoritmo de Google no afecta al método, ya que puede ser ejecutado en tiempos diferentes y desde este se puede obtener reglas actualizadas.
- Como se apreció en los trabajos previos, estos utilizan diferentes métodos y herramientas para identificar los factores que posicionen a una página web, sin embargo, ninguno propone un método sistemático que permita identificar, de una lista de factores, cuáles de estos son necesarios para implementarse en una página web de forma personalizada, es decir, recomendar a los propietarios de sitios web los factores idóneos que le sirvan para posicionarse. Esto demuestra que no es necesario aplicar todos los factores que las literaturas recomienden, sino aplicar los necesarios

para hacer SEO, lo cual guarda relación con la definición sobre SEO por parte del Equipo de Calidad de la Búsqueda de Google (2018), donde mencionan que SEO es hacer *pequeñas modificaciones* en la página web (contenido y código) y que estos cambios pueden mejorar el posicionamiento.

7.2 Trabajo futuros

Para trabajos futuros se considerará el uso de más factores de posicionamiento, su evolución y mayor número de páginas web, así como la de utilizar equipos tecnológicos que permitan realizar el rastreo de páginas web y la extracción de métricas con mayor velocidad.

REFERENCIAS BIBLIOGRÁFICAS

Aha, D., Kibler, D., & Albert, M. (1991). Instance based learning algorithms. Machine Learning, 6:37–66.

Al-Jadaan, M. (2015). An Assessment Model for SEO Techniques for Optimizing Web Visibility in Web Search Results. Tesis de Masters of Science in Software Engineering, Prince Sultan University, College of Computer and Information Sciences.

Allen, R. (13 de Abril de 2017). Search Engine Statistics 2017. Obtenido de Smart Insights: http://www.smartinsights.com/search-engine-marketing/search-engine-statistics/

Anderson, S. (04 de Octubre de 2017). SEO Tutorial For Beginners in 2017. Obtenido de https://www.hobo-web.co.uk/seo-tutorial/#what-is-seo-

Aregay, T. (2014). Ranking Factors for Web Search : Case Study in the Netherlands. Faculty of Electrical Engineering, Mathematics and Computer Science, Department of Computer Science. Netherlands: University of Twente.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, Vol 4, 40-79.

Bécares Pérez, M. (2013). Métrica de factores on-page en el posicionamiento de páginas web en los motores de búsqueda orgánicos. Tesis de Master, Universidad de Oviedo, Ingeniería y Arquitectura.

Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer Networks, 3825–3833.

Calleja Gómez, A.(2010) Minería de datos con Weka para la predicción del precio de automóviles de segunda mano. Universidad Politécnica de Valencia. Valencia

Carreras Lario, R. (2014). Cómo Clasifica Google los resultados de las búsquedas: Factores de Posicionamiento Orgánico. Tesis Doctorado, Universidad Complutense de Madrid, Facultad de Ciencias de la Información.

Carreras Lario, R (2014). Toreando a Google (Spanish Edition). Iberanálisis SL. España

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. Machine Learning, 261–283.

Clarke, A. (2017). SEO 2017: Learn search engine optimization with smart internet marketing strategies. Simple Effectiveness, 21-23.

Cómo medir los resultados de la búsqueda orgánica y de pago. (03 de Octubre de 2017). Obtenido de Ayuda de AdWords: https://support.google.com/adwords/answer/3097241?hl=es-419

Cost, S., & Salzberg, S. (1993). A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. Machine Learning, 57–78.

Craswell, N., & Hawking, D. (2009). Web Information Retrieval, in Information Retrieval: Searching in the 21st Century (eds A. Göker and J. Davies),. Chichester: City University London, UK.

Cunningham, S., Littin, J., & Witten, I. (1997). Applications of machine learning in information retrieva. University of Waikato.

Chhabra, S., Mittal, R., & Sarkar, D. (2016). Inducing factors for search engine optimization techniques: A comparative analysis. 2016 1st India International Conference on Information Processing (IICIP) (págs. 1-4). Delhi, India: IEEE.

DiSilvestro, A. (3 de Marzo de 2013). Forecasting for the Future: How to Track Google Algorithm Updates. Obtenido de https://www.searchenginejournal.com/forecasting-for-the-future-how-to-track-google-algorithm-updates/182208/

Domingos, P., & Pazzani, M. (1996). Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. Machine Learning: Proceedings of the Thirteenth International Conference on Machine Learning. Morgan Kaufmann.

Duklan, N., Mourya, D., & Bahuguna, H. (2015). Classification of search engine optimization techniques: A data mining approach. Conference: 2nd International Conference on System Modeling & Advancement in Research Trends (págs. 1-6). Moradabad, India: ,Teerthanker Mahaveer University.

Egri, G., & Bayrak, C. (2014). The Role of Search Engine Optimization on Keeping the User on the Site. Procedia Computer Science, 335-342.

Equipo de Calidad de la Búsqueda de Google. (2018). Guía de optimización en buscadores (SEO) para principiantes - Ayuda de Search Console.

El gráfico de Conocimiento. (03 de Octubre de 2017). Obtenido de Dentro deGoogle: https://www.google.com/intl/es/insidesearch/features/search/knowledge.html

Eswarawaka, R., Kudikala, S. K., Kuchi, S. C., & Verma K, V. (2017). The analysis on search engine optimization supported by six sigma methodology. 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (págs. 653-658). Bangalore, India: IEEE.

Fayyad U. M. , Piatetsky-Shapiro G. & Smyth P. (1996). From Datamining to knowledge discovery in Databases: an overview. Ai Magazine. pp. 37-54.

Florián Noriega, J. A., & Caicedo Torres, W. (2013). Categorización de texto usando técnicas de machine learning aplicado a la clasificación de reclamos

en los procesos de la Universidad Tecnológica de Bolívar. Cartagena de Indias : J. A. Florián Noriega.

Gervilla, E., Jiménez, R., Montaño, J., Sesé, A., Cajal, B., & Palmer, A. (2009). The methodology of Data Mining: An application to alcohol consumption in teenagers. Adicciones, pp. 65-80.

Gianluca, F. (30 de Septiembre de 2014). El mito de los 200 factores de posicionamiento de Google. Obtenido de MOZ: https://moz.com/blog/the-myth-of-googles-200-ranking-factors

Google. (2010). Search Engine Optimization Starter Guide. Guia. Google Inc.

Google Search Console Help. (n.d). Doorway Pages. Obtenido de https://support.google.com/webmasters/answer/2721311?hls=en&ref_topic=6001971

Grzywaczewski, A. (2010). E-Marketing Strategy for Businesses. IEEE 7th International Conference on e-Business Engineering (ICEBE) (págs. 428 – 434). Shanghai, China: IEEE.

Gudivada, V., Rao, D., & Paris, J. (2015). Understanding Search-Engine Optimization. Computer, vol. 48, no. 10, pp. 43-52.

Gupta, S., Rakesh, N., Thakral, A., & Chaudhary, D. K. (2016). Search engine optimization: Success factors. 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC) (págs. 17-21). Waknaghat, India: IEEE.

Hernández, J., Ramírez, M., & Ferri, C. (s.f.). Introducción a la minería de datos. Pearson.

HuaAAAng, W. Z. (2014). Research on Web Search Engine Optimization and its Application. Applied Mechanics and Materials, Vols. 687-691, pp. 1908-1911.

Hussien, A. (2014). Factors Affect Search Engine Optimization . IJCSNS International Journal of Computer Science and Network Security, Vol. 14 No. 9 pp. 28-33.

Illyes, G. (2016). Penguin is now part of our core algorithm. Obtenido de https://webmasters.googleblog.com/2016/09/penguin-is-now-part-of-our-core.html

Internet Live Stats. (Junio de 2018). Obtenido de www.internetlivestats.com

John, G., Kohavi, R., & Pfleger, P. (1994). Irrelevant features and the subset. Machine Learning: Proceedings of the Eleventh International. Morgan Kaufmann.

Kitchenham, B. (2004). Procedures for Perfoming Systematic Reviews. Keele University.

Kobayashi, M. & Takeda, K. (2000) Information retrieval on the web, ACM Computing Surveys, vol. 32, no. 2, pp. 144–173.

Kohavi , R., & John, G. (1996). Wrappers for feature subset selection. Artificial Intelligence, special issue on relevance, 97(1–2):273–324.

Kononenko, I., & Bratko, I. (1991). Information-based evaluation criterion for classifier's performance. Machine Learning, 6:67–80.

Krrabaj, S., Baxhaku , F., & Sadrijaj, D. (2017). Investigating search engine optimization techniques for effective ranking: A case study of an educational site. 2017 6th Mediterranean Conference on Embedded Computing (MECO) (págs. 1-4). Bar, Montenegro: IEEE.

Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. UAI'94 Proceedings of the Tenth international conference on Uncertainty in artificial intelligence, 399-406.

Langley, P., & Sage, S. (1994). Scaling to domains with irrelevant features. Computational Learning Theory and Natural Learning.

Langville & Meyer (2006). Google's PageRank and Beyond: The Science of Search Engine Ranking. Princeton University Press

Lavania, K. K., Jain, S., Gupta, M. K., Sharma, N. (2013), Google: A Case Study (Web Searching and Crawling), International Journal of Computer Theory and Engineering, Vol. 5, No. 2, April 2013, pp. 337-340.

Ledford (2008). Search Engine Optimization Bible. Wiley Publishing.USA

Lin , T.-F., & Chi , Y.-P. (2014). Application of Webpage Optimization for Clustering System on Search Engine V Google Study. International Symposium on Computer, Consumer and Control (págs. 698-701). Taichung, Taiwan: IEEE.

López Gómez, M. (Mayo de 2011). Libro SEO Posicionamiento en Buscadores (edición 3.1). Bubok Publishing S.L. Obtenido de Libro SEO, Posicionamiento En Buscadores.

Malagam, R. (2008). Worst practices in search engine optimization, Communications of the ACM. Surviving the data deluge, 12.

Martínez, N.(2016). Predicción y Análisis de los Retrasos en los
Vuelos. Estudio del Aeropuerto de Arizona (E.E.U.U.) Memoria del Proyecto de Fin de Grado. Universidad Autónoma de Barcelona

Martínez Álvarez, c.a.(2012) Aplicación de técnicas de minería de datos para

mejorar el proceso de control de gestión en Entel. Tesis Mag. Santiago de Chile. 111p.

Mitchell, T. (1990). Machine Learning [Aprendizaje automático]. Annual Review of Computer Science, Vol. 4:417-433 .

Mitchell, T. (1997). Machine Learning (McGraw-Hill International Editions Computer Science Series) 1st Edition. McGraw-Hill, Inc.

Moráguez, M., & Cancio, L. (2014). Propuesta de factores a considerar en el posicionamiento de los sitios web de salud . Revista Internacional de Gestión del Conocimiento y la Tecnología, 10-30.

Morato, J., Sánchez-Cuadrado, S., Moreno, V., & Moreiro, J. (2013). Evolución de los factores de posicionamiento web y adaptación de las herramientas de optimización. Revista Española de Documentación Científica.

Nasomyon, T., & Wisitpongphan, N. (2014). A Study on The Relationship between Search Engine Optimization Factors and Rank on Google Search Result Page. Advanced Materials Research , 1462-1466 .

Ng, K., & Liu, H. (2001). Customer Retention via Data Mining. Artificial Intelligence Review, pp. 569-590.

Ñaupas Caraza, C. M. (2016). Minería de datos aplicada a la detección de fraude electrónico en entidades bancarias.

Ochoa, E. (2012). An Analysis of the Application of Selected Search Engine Optimization (SEO) Techniques and Their Effectiveness on Google's Search Ranking Algorithm. California State: Dissertation, California State University.

Página de resultados del buscador - Wikipedia. (03 de Octubre de 2017). Obtenido de Wikipedia: https://es.wikipedia.org/wiki/P%C3%A1gina_de_resultados_del_buscador

Palabra clave - Wikipedia. (03 de Octubre de 2017). Obtenido de Wikipedia: https://es.wikipedia.org/wiki/Palabra_clave

Page, L. (2001). Patente nº US 6285999 B1. EEUU.

PageRank - Wikipedia. (30 de Diciembre de 2018). Obtenido de Wikipedia: https://es.wikipedia.org/wiki/PageRank

Patel, N. (2014). The seven commmandments of internal linking that will improve content marketing SEO. Obtenido de https://blog.kissmetrics.com/commandments-of-internal-linking/

Pérez-Montoro, M., & Codina, L. (2017). Navigation design and seo for contentintensive websites: A Guide for an Efficient. Chandos Publishing.

Prabha, S., Duraiswamy, K., & Indhumathi, J. (2014). Comparative Analysis of Different Page Ranking Algorithms. World Academy of Science, Engineering and Technology, International Science Index 92, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 8(8), 1546 - 1554.

Quinlan, J. (1986). Induction of Decision Trees . Machine Learning, 81-106.

Quinlan, J. (1993). C4.5: Programs for machine learning. Morgan Kaufmann, 235-240.

Rayhan, H. (2013). Improve Webseite Rank Using Search Engine Optimization(SEO). Faculty of Computer & Information Al-Madinah International University. University of Jordan.

Rehman, K., & Khan, M. (2013). The Foremost Guidelines for Achieving Higher Ranking in Search Results through Search Engine Optimization.

International Journal of Advanced Science and Technology, Vol.52, pp.101-110.

Rokach, L., & Maimon, O. (2008). Data Mining with Decision Trees: Theory and Applications. World Scientific Publishing.

Roldán, R. (26 de Junio de 2017). Social Media Líderes. Obtenido de El proceso SEO en 5 Fases: http://socialmedialideres.com.ve/proceso-seo-en-5-fases/

Sandhya , D., Thakare, V., & Butey, P. (2016). SEO Techniques for various Applications - A Comparative Analyses and Evaluation . IOSR Journal of Computer Engineering (IOSR-JCE) , pp 20-24.

SantaMaria Ruiz, W. (2010). Modelo de Detección de fraude basado en el Descubrimiento de reglas de clasificación extraídas de una red neuronal. Tesis de Magister, Universidad Nacional de Colombia. , Ingeniería de Software s y Computación.

Sarika, R., & Sharma, S. (2014). An Analytical Study of the Search Engine Optimization Techniques for Information Retrieval Systems. International Journal of Research in Computer Science and Management (Vol.2, No. 1), 4 - 13.

Search Console Help. (2017). Obtenido de File types indexable by Google: https://support.google.com/webmasters/answer/35287?hl=en

SEO.es PowerTools. (2016). Obtenido de Historia de los Algoritmos más importantes de Google: https://www.seo.es/que-es/algoritmos-de-google

Shubham, S., Shubham, S., & Shweta, K. (2015). Search Engine Optimization with Page Rank. International Journal of Innovations & Advancement in Computer Science(JIACS), Volume 4, 177-185.

Silva, N., & Aguiar, A. (2014). Optimização de Sítios Web para Motores de Busca Um Estudo Empírico. Information Systems and Technologies (CISTI), 2014 9th Iberian Conference on. Barcelona, Spain: IEEE.

Singhal , A. (2011). More guidance on building high–quality sites. Posted 06 May. Accessed 19 December 2016. Obtenido de https://webmasters.googleblog.com/2011/05/more-guidance-on-building-high-quality.html

Smit, M. (1993). Neuronal Networks for Statistical Modeling. New York,: John Wiley & Sons.

Su, A.-J., Hu, Y., Kuzmanovic, A., & Koh, C.-K. (2014). How to improve your search engine ranking: Myths and reality. ACM Transactions on the Web, Vol. 8, No. 2.

Sylvain Sagot, E., & Fougèr, A.-J. (2016). A multi-agent approach for building a fuzzy decision support system to assist the SEO process . 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (págs. 004001-004006). Budapest, Hungary: IEEE.

Tejada, V. (3 de Julio de 2017). Vanessa Tejada. Obtenido de Google Búho, el Nuevo Algoritmo de Google: http://vanessatejada.com/sobre-google/google-buho-el-nuevo-algoritmo-de-google/

Themistoklis, M., & Symeonidis , A. (2015). Identifying valid search engine ranking factors in a Web 2.0 and Web 3.0 context for building efficient SEO mechanism. Engineering Applications of Artificial Intelligence, Volume 41, Pages 75-91.

Tober, M., Hennig, L., & Furch, D. (2014). SEO Ranking Factors and Rank Correlations 2014. Whitepaper Searchmetrics.

Two Crows Corporation. (1999). Introduction to Data Mining and Knowledge Discovery. 3° ed. Two Crows Corporation.

Velásquez, J., & Palade, V. (2008). Adaptive Web Sites: A Knowledge Extraction from Web Data Approach. IOS Press, 1-272.

White, B. (2015). An update on doorway pages. Obtenido de https://webmasters.googleblog.com/2015/03/an-update-on-doorway-pages.html

Wu, X. (2008). Top 10 algorithms in data mining. Springer-Verlag.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, pp. 645-678.

Xu, R., & Wunsch, D. (2009). Clustering. Wiley Publishing